Robust Consistent Video Depth Estimation
Supplementary Material

In this supplementary document, we present additional implementation details, quantitative evaluation, and visual results to complement the main paper [2]

1. Visual Results of Depth and Pose Estimation

We provide the visual comparisons with the state-of-the-art depth estimation algorithms in the accompanying videos. Specifically, we show the following three videos:

1. **Cellphone videos**: We show visual comparisons of our results with Consistent Video Depth (CVD) [4], COLMAP [6], and DeepV2D [9].
2. **Sintel dataset**: We use sample videos from the Sintel dataset to validate the two core technical contributions of our work (1) flexible depth deformation for pose optimization and (2) geometry-aware depth filtering.
3. **DAVIS dataset**: To demonstrate the robustness of our approach, we present the results of *all* 90 videos from the DAVIS 2017 train and validation dataset.

We will release the source code and the results (dense depth maps and camera poses).

2. Runtime Analysis

We provide a runtime analysis and profiling of our method. Given a long input video, we test our method with a varying length of input video frames (first K frames from the video). The original video resolution: 1920×1080; downscaled resolution for optical flow estimation [10]: 1024×576; downscaled resolution for initial depth estimation [5]: 384×224). We use two NVIDIA Tesla V100 GPUs. Note that our depth filtering code is *single-threaded* and would significantly benefit from multi-threading in practice. Figure 1 and Table 1 demonstrate the detailed analysis.

Discussion. The performance of our method is far from real-time, but compares favorably to many of the baseline methods we have tested: COLMAP-dense [7] runtime is very variable, but generally considerably slower than our method, on the order of hours or even days in our experiments. CVD [4] typically runs for multiple hours on a short video clip. DeepV2D [9] is faster (e.g. a few minutes for a 50-frame sequence); however, the quality is considerably lower than ours, and it quickly runs out of memory for longer sequences (the maximum were 75 frames on a NVIDIA 2080Ti GPU with 11GB memory). We are not aware of any real-time dense video depth estimation methods that provide comparable quality to ours. Most SLAM methods, for example, only provide sparse or semi-dense depth.

Table 1. Runtime analysis of videos of varying length, broken down by algorithm stages. Times are reported in seconds.

Video Length in Frames	Optical Flow [10]	Depth Estimation [5]	Pose Optimization	Geometry-aware Depth Filtering	Total
50	252.0s	9.2s	36.6s	75.8s	373.6s
100	515.7s	16.0s	54.5s	172.0s	758.3s
200	1,268.9s	30.9s	94.8s	334.9s	1,729.5s
400	2,611.4s	58.3s	175.2s	704.1s	3,549.0s

3. Evaluation on ScanNet

Table 2 provides a quantitative evaluation on the ScanNet dataset.

4. Additional Quantitative Evaluation on the Sintel Dataset

Depth evaluation partial results. Figure 2 shows a quantitative comparison with methods that produce *partial* depth maps (i.e., depth is not reported for all pixels; e.g., COLMAP) or do not succeed in some of the sequences. To quantitatively

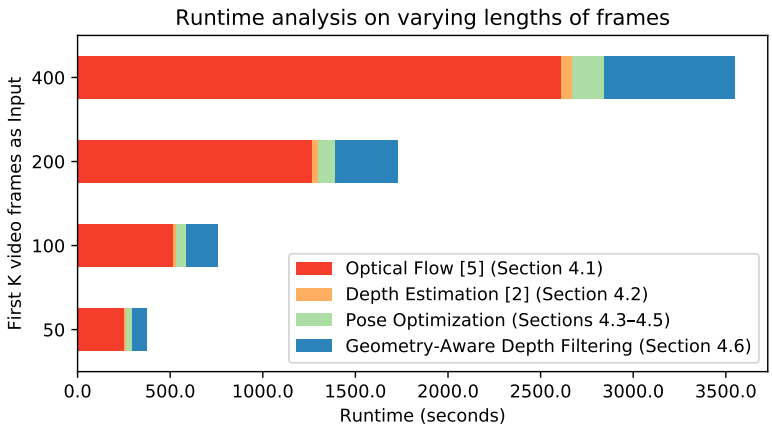
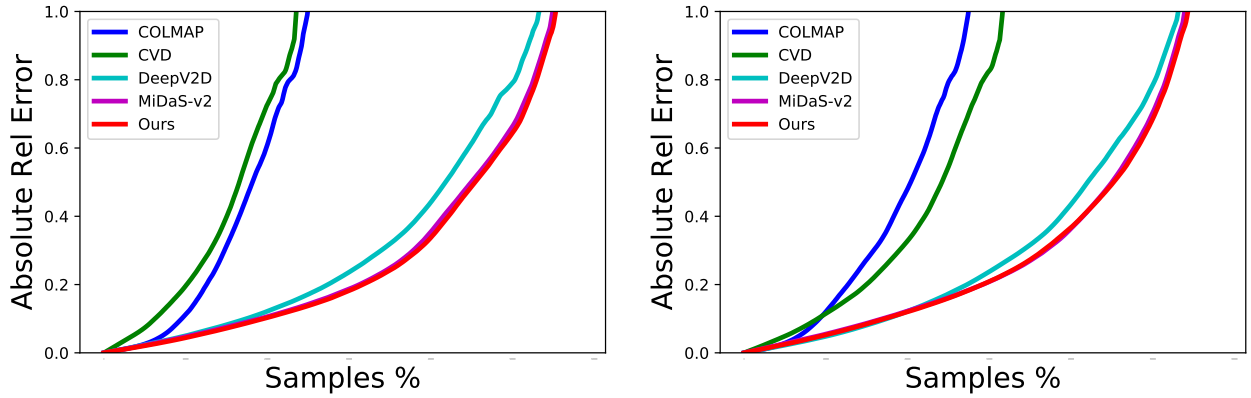


Figure 1. **Runtime analysis of the runtime of different stages in our proposed approach.**

Table 2. Quantitative comparison on the ScanNet dataset [1] using the test split provided by Tang and Tan [8].

	Error metric ↓				
	Abs Rel	Sq Rel	RMSE	RMSE log	Sc Inv
DeMoN [11]	0.231	0.520	0.761	0.289	0.284
BA-Net [8]	0.161	0.092	0.346	0.214	0.184
DeepV2D (NYU) [9]	0.080	0.018	0.223	0.109	0.105
DeepV2D (ScanNet) [9]	**0.057**	**0.010**	**0.168**	**0.080**	**0.077**
MiDaS-v2 [3]	0.208	0.318	0.742	0.246	0.239
CVD [4]	0.073	0.037	0.217	0.105	0.103
Ours	0.180	0.188	0.603	0.218	0.217

evaluate the performance for methods with partial results, we store all the pixel-wise error metrics across all the pixels, frames, and videos and plot the sorted values. The plot here capture both the accuracy as well as the completeness of each method.



(a) Abs Rel - Sintel Clean (b) Abs Rel - Sintel Final

Figure 2. Comparisons of Absolute Rel Error on Sintel benchmark. (*Clean* and *Final* categories respectively). All the per-frame errors are stored and sorted for plotting the distributions. Note that COLMAP and CVD (which relies on COLMAP) fail on many Sintel sequences, resulting in partial data points.

Per-sequence evaluation. We provide per-sequence quantitative evaluation for depth and pose on the Sintel dataset in Table 3, Table 4, and Table 5.

5. Additional Discussion (from Rebuttal)

Can the inaccurate depth itself be directly optimized, instead of introducing an extra deformation model? What are the advantages of using a deformation model? (1) It will be too unconstrained. In CVD, with fixed camera poses, optimizing depth values directly would reduce the problem to naïve multi-view stereo triangulation (which would not produce good results). In our case, where camera poses are part of the optimization variables, direct depth value optimization would not converge to a good solution, since the problem is non-linear and sensitive to initialization. Overcoming the sensitivity to initialization would require good regularization. CVD achieves it by fine-tuning CNN weights shared for all frames, rather than optimizing the depth values themselves. In our case, we achieve the regularization by using smooth depth deformation functions, which enforce a similarity (up to scale) to the original depth produced by the pretrained depth network. (2) The global optimization (for all frames) problem is too large. The smooth deformation functions have fewer parameters, which makes global optimization feasible

Is smooth depth deformation an ideal solution? Ideally we would use a *piece-wise* smooth deformation that is controlled at the pixel level. However, we are currently bound by the memory and compute demands of the global deformation optimization. Please refer to our response to the question about limitations above. We believe this is a great avenue for future work.

Table 3. **Per-sequence quantitative evaluations of depth and pose on the MPI Sintel benchmark** (*Sintel Final*).

	Depth - Error metric↓				Depth - Accuracy metric↑			Pose - Error metric↓		
Method	**Abs Rel**	**Sq Rel**	**RMSE**	**log RMSE**	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	**ATE (m)**↓	**RPE Trans (m)**↓	**RPE Rot (deg)**↓
					alley_1					
DeepV2D [9]	1.274	4.724	4.145	0.908	0.672	0.738	0.752	0.103	0.043	0.023
Ours - Single-scale pose (aligned MiDaS)	0.511	0.969	3.146	0.512	0.602	0.722	0.769	0.112	0.071	0.023
Ours - Single-scale pose + depth fine-tuning	1.077	3.499	3.784	0.832	0.715	0.751	0.753	0.033	0.012	0.006
Ours - Single-scale pose + depth filter	0.511	0.956	3.135	0.512	0.600	0.731	0.770	0.112	0.071	0.023
Ours - Flexible pose	0.532	1.170	3.252	0.522	0.621	0.740	0.809	0.026	0.015	0.002
Ours - Flexible pose + depth fine-tuning	0.874	2.402	3.807	0.751	0.707	0.751	0.753	0.026	0.010	0.003
Ours - Flexible pose + depth filter	0.531	1.157	3.250	0.521	0.625	0.740	0.808	0.026	0.015	0.002
					alley_2					
DeepV2D [9]	0.317	2.324	7.561	0.560	0.485	0.736	0.825	0.807	0.382	0.158
Ours - Single-scale pose (aligned MiDaS)	0.263	1.377	5.988	0.382	0.556	0.806	0.918	0.327	0.082	0.021
Ours - Single-scale pose + depth fine-tuning	0.379	2.423	7.646	0.594	0.387	0.624	0.825	0.360	0.073	0.011
Ours - Single-scale pose + depth filter	0.261	1.370	5.996	0.381	0.557	0.810	0.917	0.327	0.082	0.021
Ours - Flexible pose	0.244	1.332	5.987	0.370	0.562	0.859	0.918	0.320	0.072	0.009
Ours - Flexible pose + depth fine-tuning	0.292	1.820	6.878	0.465	0.480	0.787	0.893	0.324	0.073	0.009
Ours - Flexible pose + depth filter	0.244	1.332	5.990	0.370	0.563	0.860	0.918	0.320	0.072	0.009
					ambush_2					
DeepV2D [9]	0.716	8.314	20.149	1.764	0.340	0.460	0.570	1.731	3.314	0.680
Ours - Single-scale pose (aligned MiDaS)	0.659	7.949	19.914	1.528	0.365	0.497	0.611	0.348	2.426	0.140
Ours - Single-scale pose + depth fine-tuning	0.702	8.303	20.140	1.747	0.349	0.464	0.581	0.311	2.434	0.132
Ours - Single-scale pose + depth filter	0.656	7.947	19.921	1.532	0.365	0.496	0.613	0.358	2.424	0.142
Ours - Flexible pose	0.657	7.958	19.923	1.532	0.361	0.497	0.610	0.348	2.468	0.112
Ours - Flexible pose + depth fine-tuning	0.686	8.125	20.049	1.648	0.351	0.467	0.582	0.385	2.466	0.123
Ours - Flexible pose + depth filter	0.656	7.953	19.927	1.533	0.360	0.496	0.611	0.348	2.468	0.112
					ambush_4					
DeepV2D [9]	0.781	2.307	2.510	0.758	0.357	0.479	0.649	1.629	1.638	1.011
Ours - Single-scale pose (aligned MiDaS)	0.610	1.566	2.483	0.651	0.335	0.510	0.697	0.240	0.285	0.071
Ours - Single-scale pose + depth fine-tuning	0.688	1.883	2.436	0.713	0.371	0.478	0.685	0.124	0.226	0.030
Ours - Single-scale pose + depth filter	0.593	1.463	2.447	0.646	0.342	0.515	0.709	0.242	0.285	0.079
Ours - Flexible pose	0.605	1.462	2.446	0.655	0.320	0.520	0.706	0.199	0.244	0.050
Ours - Flexible pose + depth fine-tuning	0.625	1.585	2.434	0.688	0.373	0.512	0.679	0.188	0.236	0.038
Ours - Flexible pose + depth filter	0.601	1.442	2.436	0.654	0.323	0.523	0.708	0.199	0.244	0.050
					ambush_5					
DeepV2D [9]	0.416	7.732	19.769	1.286	0.363	0.642	0.732	1.294	1.240	0.962
Ours - Single-scale pose (aligned MiDaS)	0.335	7.329	19.421	1.158	0.548	0.727	0.756	0.277	0.333	0.038
Ours - Single-scale pose + depth fine-tuning	0.354	7.807	19.907	1.332	0.517	0.718	0.745	0.165	0.304	0.031
Ours - Single-scale pose + depth filter	0.330	7.331	19.428	1.157	0.561	0.732	0.757	0.291	0.329	0.046
Ours - Flexible pose	0.345	7.388	19.483	1.178	0.518	0.708	0.753	0.199	0.298	0.030
Ours - Flexible pose + depth fine-tuning	0.397	7.869	19.903	1.346	0.413	0.673	0.746	0.200	0.312	0.027
Ours - Flexible pose + depth filter	0.339	7.390	19.490	1.176	0.530	0.715	0.754	0.199	0.298	0.030
					ambush_6					
DeepV2D [9]	0.615	1.071	2.316	0.681	0.385	0.595	0.703	2.603	2.576	1.058
Ours - Single-scale pose (aligned MiDaS)	0.315	0.725	2.342	0.458	0.543	0.749	0.880	0.292	1.266	0.071
Ours - Single-scale pose + depth fine-tuning	0.535	0.919	2.278	0.643	0.399	0.609	0.722	0.230	1.287	0.068
Ours - Single-scale pose + depth filter	0.302	0.689	2.320	0.449	0.566	0.754	0.888	0.236	1.257	0.079
Ours - Flexible pose	0.323	0.742	2.367	0.471	0.541	0.751	0.879	0.197	1.274	0.062
Ours - Flexible pose + depth fine-tuning	0.433	0.710	2.191	0.563	0.433	0.650	0.784	0.287	1.279	0.072
Ours - Flexible pose + depth filter	0.317	0.724	2.358	0.466	0.551	0.755	0.880	0.197	1.274	0.062
					ambush_7					
DeepV2D [9]	0.064	0.005	0.053	0.089	0.973	0.999	1.000	0.632	0.290	0.193
Ours - Single-scale pose (aligned MiDaS)	0.125	0.014	0.085	0.148	0.847	0.995	1.000	0.208	0.053	0.035
Ours - Single-scale pose + depth fine-tuning	0.051	0.003	0.040	0.067	0.997	1.000	1.000	0.120	0.019	0.014
Ours - Single-scale pose + depth filter	0.120	0.012	0.080	0.142	0.865	0.999	1.000	0.160	0.049	0.031
Ours - Flexible pose	0.108	0.009	0.071	0.136	0.896	0.999	1.000	0.091	0.022	0.011
Ours - Flexible pose + depth fine-tuning	0.081	0.005	0.056	0.104	0.983	1.000	1.000	0.202	0.031	0.015
Ours - Flexible pose + depth filter	0.106	0.009	0.069	0.132	0.912	0.999	1.000	0.091	0.022	0.011
					bamboo_1					
DeepV2D [9]	0.513	2.243	2.192	0.519	0.696	0.788	0.842	0.729	0.259	0.141
Ours - Single-scale pose (aligned MiDaS)	0.539	2.368	2.351	0.532	0.686	0.781	0.834	0.133	0.075	0.026
Ours - Single-scale pose + depth fine-tuning	0.493	2.051	2.124	0.507	0.684	0.793	0.844	0.079	0.019	0.013
Ours - Single-scale pose + depth filter	0.536	2.362	2.339	0.531	0.692	0.781	0.834	0.133	0.075	0.026
Ours - Flexible pose	0.477	1.955	2.085	0.501	0.707	0.791	0.845	0.121	0.018	0.004
Ours - Flexible pose + depth fine-tuning	0.494	2.070	2.123	0.509	0.698	0.790	0.844	0.097	0.016	0.006
Ours - Flexible pose + depth filter	0.478	1.960	2.088	0.501	0.707	0.790	0.845	0.121	0.018	0.004
					bamboo_2					
DeepV2D [9]	0.518	1.792	4.190	0.636	0.350	0.602	0.766	0.431	0.125	0.083
Ours - Single-scale pose (aligned MiDaS)	0.428	1.433	4.041	0.572	0.400	0.653	0.806	0.110	0.067	0.024
Ours - Single-scale pose + depth fine-tuning	0.482	1.622	4.255	0.627	0.294	0.612	0.778	0.018	0.020	0.011
Ours - Single-scale pose + depth filter	0.426	1.426	4.040	0.570	0.408	0.654	0.809	0.110	0.067	0.024
Ours - Flexible pose	0.415	1.348	3.975	0.552	0.386	0.688	0.821	0.040	0.014	0.003
Ours - Flexible pose + depth fine-tuning	0.494	1.656	4.240	0.629	0.303	0.596	0.775	0.015	0.016	0.006
Ours - Flexible pose + depth filter	0.415	1.348	3.977	0.552	0.387	0.688	0.821	0.040	0.014	0.003

Table 4. **Per-sequence quantitative evaluations of depth and pose on the MPI Sintel benchmark** (*Sintel Final*).

	Depth - Error metric↓				Depth - Accuracy metric↑			Pose - Error metric↓		
Method	**Abs Rel**	**Sq Rel**	**RMSE**	**log RMSE**	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	**ATE (m)**↓	**RPE Trans (m)**↓	**RPE Rot (deg)**↓
					bandage_1					
DeepV2D [9]	0.214	0.084	0.334	0.329	0.591	0.845	0.922	0.762	0.425	0.255
Ours - Single-scale pose (aligned MiDaS)	0.864	3.635	1.714	0.674	0.624	0.712	0.776	0.077	0.034	0.002
Ours - Single-scale pose + depth fine-tuning	0.146	0.051	0.272	0.241	0.765	0.883	0.990	0.083	0.016	0.007
Ours - Single-scale pose + depth filter	0.857	3.418	1.694	0.675	0.625	0.709	0.775	0.077	0.034	0.002
Ours - Flexible pose	0.850	3.352	1.676	0.673	0.624	0.714	0.779	0.070	0.026	0.002
Ours - Flexible pose + depth fine-tuning	0.125	0.040	0.240	0.208	0.825	0.912	0.997	0.077	0.011	0.005
Ours - Flexible pose + depth filter	0.845	3.210	1.661	0.674	0.625	0.711	0.780	0.070	0.026	0.002
					bandage_2					
DeepV2D [9]	0.438	0.231	0.474	0.469	0.260	0.627	0.872	0.062	0.019	0.007
Ours - Single-scale pose (aligned MiDaS)	0.550	1.135	1.255	0.527	0.400	0.631	0.796	0.103	0.042	0.010
Ours - Single-scale pose + depth fine-tuning	0.418	0.184	0.412	0.424	0.370	0.612	0.909	0.034	0.018	0.006
Ours - Single-scale pose + depth filter	0.542	1.066	1.229	0.522	0.399	0.631	0.799	0.106	0.045	0.010
Ours - Flexible pose	0.492	0.668	0.960	0.513	0.346	0.591	0.801	0.082	0.022	0.003
Ours - Flexible pose + depth fine-tuning	0.470	0.241	0.487	0.484	0.229	0.559	0.864	0.045	0.020	0.004
Ours - Flexible pose + depth filter	0.488	0.644	0.940	0.511	0.347	0.584	0.805	0.082	0.022	0.003
					cave_2					
DeepV2D [9]	0.972	16.038	18.036	0.930	0.281	0.435	0.545	1.589	3.550	1.114
Ours - Single-scale pose (aligned MiDaS)	0.700	10.207	15.807	0.730	0.312	0.524	0.651	0.941	3.034	0.081
Ours - Single-scale pose + depth fine-tuning	0.904	14.897	17.791	0.894	0.301	0.448	0.569	0.879	3.036	0.087
Ours - Single-scale pose + depth filter	0.688	10.371	15.901	0.725	0.317	0.530	0.654	1.113	3.041	0.081
Ours - Flexible pose	0.689	9.851	15.988	0.735	0.296	0.522	0.658	0.968	3.027	0.083
Ours - Flexible pose + depth fine-tuning	0.810	11.986	17.659	0.877	0.324	0.473	0.573	0.789	3.020	0.082
Ours - Flexible pose + depth filter	0.683	9.686	15.969	0.732	0.299	0.528	0.659	0.968	3.027	0.083
					cave_4					
DeepV2D [9]	0.444	2.244	5.513	0.631	0.355	0.595	0.733	0.936	1.530	0.840
Ours - Single-scale pose (aligned MiDaS)	0.365	1.662	4.868	0.524	0.411	0.660	0.805	0.188	0.977	0.044
Ours - Single-scale pose + depth fine-tuning	0.406	2.041	5.403	0.602	0.375	0.631	0.759	0.177	0.962	0.030
Ours - Single-scale pose + depth filter	0.359	1.632	4.855	0.518	0.413	0.666	0.807	0.184	0.977	0.045
Ours - Flexible pose	0.356	1.684	4.965	0.529	0.426	0.673	0.797	0.205	0.960	0.030
Ours - Flexible pose + depth fine-tuning	0.425	2.135	5.449	0.613	0.372	0.615	0.748	0.167	0.948	0.047
Ours - Flexible pose + depth filter	0.353	1.669	4.951	0.526	0.429	0.676	0.797	0.205	0.960	0.030
					market_2					
DeepV2D [9]	0.613	10.914	18.792	0.648	0.307	0.499	0.662	0.433	0.233	0.169
Ours - Single-scale pose (aligned MiDaS)	0.254	3.281	12.238	0.336	0.564	0.815	0.954	0.089	0.051	0.017
Ours - Single-scale pose + depth fine-tuning	0.459	7.052	16.473	0.530	0.388	0.612	0.735	0.042	0.018	0.008
Ours - Single-scale pose + depth filter	0.252	3.235	12.204	0.333	0.565	0.817	0.957	0.089	0.051	0.017
Ours - Flexible pose	0.345	4.641	13.972	0.413	0.388	0.720	0.887	0.045	0.020	0.003
Ours - Flexible pose + depth fine-tuning	0.507	8.346	18.137	0.589	0.309	0.530	0.699	0.054	0.016	0.006
Ours - Flexible pose + depth filter	0.344	4.598	13.963	0.412	0.384	0.715	0.889	0.045	0.020	0.003
					market_5					
DeepV2D [9]	0.780	2.505	4.181	0.846	0.246	0.435	0.560	2.464	1.406	1.048
Ours - Single-scale pose (aligned MiDaS)	0.418	1.253	3.428	0.565	0.385	0.634	0.782	1.974	0.316	0.053
Ours - Single-scale pose + depth fine-tuning	0.666	2.062	4.053	0.776	0.268	0.467	0.604	1.982	0.277	0.036
Ours - Single-scale pose + depth filter	0.432	1.247	3.409	0.559	0.378	0.616	0.775	1.768	0.326	0.056
Ours - Flexible pose	0.417	1.230	3.375	0.560	0.408	0.625	0.781	1.986	0.308	0.032
Ours - Flexible pose + depth fine-tuning	0.517	1.784	4.026	0.734	0.306	0.551	0.690	1.943	0.281	0.032
Ours - Flexible pose + depth filter	0.432	1.224	3.352	0.552	0.392	0.608	0.776	1.986	0.308	0.032
					market_6					
DeepV2D [9]	0.951	6.073	12.243	1.131	0.184	0.293	0.409	1.425	1.374	0.993
Ours - Single-scale pose (aligned MiDaS)	0.655	4.255	10.912	0.850	0.234	0.377	0.541	0.715	0.359	0.033
Ours - Single-scale pose + depth fine-tuning	0.757	5.154	11.906	1.003	0.206	0.324	0.467	0.717	0.344	0.024
Ours - Single-scale pose + depth filter	0.649	4.230	10.916	0.846	0.236	0.378	0.542	0.715	0.359	0.033
Ours - Flexible pose	0.590	3.976	10.708	0.804	0.241	0.400	0.634	0.695	0.349	0.016
Ours - Flexible pose + depth fine-tuning	0.620	4.822	11.863	0.952	0.227	0.378	0.570	0.612	0.352	0.026
Ours - Flexible pose + depth filter	0.587	3.965	10.711	0.803	0.242	0.399	0.639	0.695	0.349	0.016
					shaman_2					
DeepV2D [9]	0.326	0.639	1.504	0.715	0.549	0.568	0.570	0.016	0.002	0.002
Ours - Single-scale pose (aligned MiDaS)	0.152	0.188	0.813	0.362	0.723	0.858	0.946	0.074	0.025	0.006
Ours - Single-scale pose + depth fine-tuning	0.349	0.703	1.568	0.783	0.561	0.568	0.569	0.068	0.014	0.005
Ours - Single-scale pose + depth filter	0.151	0.186	0.809	0.360	0.724	0.863	0.948	0.074	0.025	0.006
Ours - Flexible pose	0.153	0.187	0.812	0.361	0.723	0.858	0.947	0.062	0.014	0.003
Ours - Flexible pose + depth fine-tuning	0.189	0.247	0.957	0.386	0.612	0.833	0.913	0.047	0.006	0.002
Ours - Flexible pose + depth filter	0.152	0.186	0.809	0.359	0.724	0.862	0.948	0.062	0.014	0.003
					shaman_3					
DeepV2D [9]	0.210	0.028	0.113	0.229	0.614	0.970	1.000	0.417	0.185	0.102
Ours - Single-scale pose (aligned MiDaS)	0.194	0.038	0.137	0.220	0.681	0.949	0.999	0.086	0.055	0.014
Ours - Single-scale pose + depth fine-tuning	0.142	0.013	0.082	0.161	0.846	1.000	1.000	0.040	0.011	0.004
Ours - Single-scale pose + depth filter	0.190	0.036	0.135	0.216	0.689	0.957	0.999	0.086	0.055	0.014
Ours - Flexible pose	0.199	0.039	0.143	0.222	0.674	0.963	0.996	0.054	0.019	0.003
Ours - Flexible pose + depth fine-tuning	0.125	0.011	0.072	0.147	0.853	1.000	1.000	0.040	0.010	0.003
Ours - Flexible pose + depth filter	0.197	0.038	0.142	0.221	0.680	0.959	0.999	0.054	0.019	0.003

Table 5. **Per-sequence quantitative evaluations of depth and pose on the MPI Sintel benchmark** (*Sintel Final*).

	Depth - Error metric↓				Depth - Accuracy metric↑			Pose - Error metric↓		
Method	**Abs Rel**	**Sq Rel**	**RMSE**	**log RMSE**	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	**ATE (m)**↓	**RPE Trans (m)**↓	**RPE Rot (deg)**↓
					sleeping_1					
DeepV2D [9]	0.060	0.005	0.059	0.081	0.980	1.000	1.000	0.011	0.009	0.008
Ours - Single-scale pose (aligned MiDaS)	0.129	0.027	0.141	0.154	0.907	0.983	0.996	0.091	0.033	0.019
Ours - Single-scale pose + depth fine-tuning	0.059	0.004	0.053	0.073	0.999	1.000	1.000	0.080	0.011	0.007
Ours - Single-scale pose + depth filter	0.128	0.026	0.139	0.152	0.914	0.983	0.996	0.091	0.033	0.019
Ours - Flexible pose	0.135	0.033	0.155	0.161	0.911	0.980	0.990	0.090	0.011	0.006
Ours - Flexible pose + depth fine-tuning	0.056	0.004	0.053	0.074	0.995	1.000	1.000	0.075	0.008	0.003
Ours - Flexible pose + depth filter	0.134	0.033	0.154	0.161	0.913	0.980	0.990	0.090	0.011	0.006
					sleeping_2					
DeepV2D [9]	0.219	0.119	0.495	0.238	0.525	0.985	1.000	0.134	0.010	0.008
Ours - Single-scale pose (aligned MiDaS)	0.279	0.193	0.602	0.288	0.400	0.913	1.000	0.161	0.049	0.024
Ours - Single-scale pose + depth fine-tuning	0.280	0.193	0.590	0.296	0.462	0.870	1.000	0.169	0.016	0.004
Ours - Single-scale pose + depth filter	0.278	0.191	0.599	0.287	0.396	0.913	1.000	0.161	0.049	0.024
Ours - Flexible pose	0.280	0.193	0.604	0.289	0.391	0.923	0.999	0.179	0.017	0.004
Ours - Flexible pose + depth fine-tuning	0.303	0.219	0.642	0.319	0.358	0.839	1.000	0.169	0.015	0.003
Ours - Flexible pose + depth filter	0.280	0.193	0.603	0.289	0.388	0.924	0.999	0.179	0.017	0.004
					temple_2					
DeepV2D [9]	0.653	5.760	10.404	0.603	0.558	0.838	0.879	2.638	0.965	0.740
Ours - Single-scale pose (aligned MiDaS)	0.614	5.168	10.410	0.582	0.502	0.755	0.884	2.227	0.268	0.017
Ours - Single-scale pose + depth fine-tuning	0.609	5.681	10.122	0.569	0.726	0.871	0.890	2.245	0.259	0.022
Ours - Single-scale pose + depth filter	0.618	5.175	10.385	0.581	0.502	0.757	0.888	2.227	0.268	0.017
Ours - Flexible pose	0.632	5.870	10.173	0.578	0.597	0.831	0.885	2.323	0.265	0.015
Ours - Flexible pose + depth fine-tuning	0.638	6.483	9.990	0.563	0.755	0.881	0.893	2.286	0.259	0.019
Ours - Flexible pose + depth filter	0.632	5.827	10.152	0.577	0.599	0.835	0.885	2.323	0.265	0.015
					temple_3					
DeepV2D [9]	0.481	4.485	7.312	0.686	0.631	0.689	0.732	2.275	2.139	1.603
Ours - Single-scale pose (aligned MiDaS)	0.384	3.315	6.790	0.544	0.606	0.729	0.807	1.111	0.532	0.099
Ours - Single-scale pose + depth fine-tuning	0.450	4.174	7.225	0.650	0.623	0.703	0.770	0.723	0.332	0.090
Ours - Single-scale pose + depth filter	0.377	3.188	6.720	0.541	0.611	0.729	0.811	1.111	0.532	0.099
Ours - Flexible pose	0.410	3.430	6.807	0.554	0.570	0.715	0.810	0.888	0.412	0.056
Ours - Flexible pose + depth fine-tuning	0.466	4.613	7.341	0.661	0.635	0.703	0.762	0.565	0.345	0.099
Ours - Flexible pose + depth filter	0.399	3.227	6.714	0.548	0.584	0.719	0.816	0.888	0.412	0.056

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2

[2] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *CVPR*, 2021. 1

[3] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019. 2

[4] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM TOG (Proc. SIGGRAPH)*, 39(4), 2020. 1, 2

[5] René Ranftl, Katrin Lasinger, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 1

[6] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[7] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1

[8] Chengzhou Tang and Ping Tan. BA-Net: Dense bundle adjustment network. In *ICLR*, 2019. 2

[9] Zachary Teed and Jia Deng. DeepV2D: Video to depth with differentiable structure from motion. In *ICLR*, 2020. 1, 2, 4, 5, 6

[10] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 1

[11] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *CVPR*, 2017. 2