DepthSync: Diffusion Guidance-Based Depth Synchronization for Scale- and Geometry-Consistent Video Depth Estimation

Yue-Jiang Dong[1], Wang Zhao[2], Jiale Xu[2], Ying Shan[2], Song-Hai Zhang[1]

[1]Tsinghua University [2]ARC Lab, Tencent PCG

Motivation

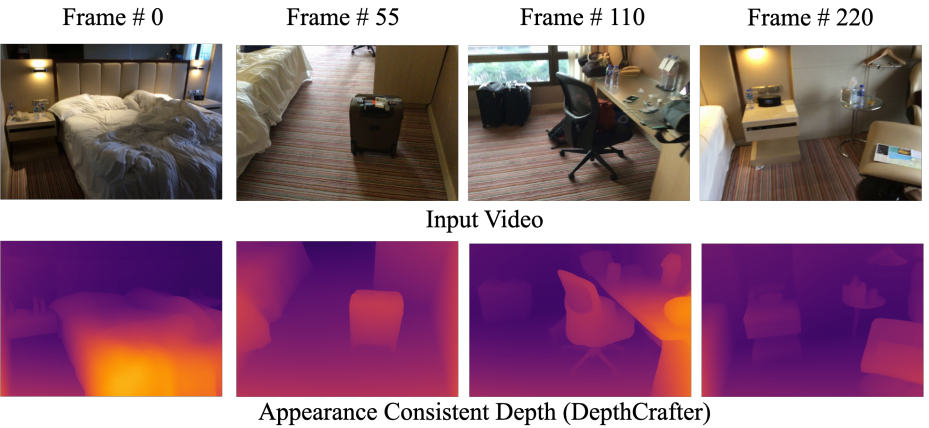
Two Types of Inconsistencies in current video diffusion-based depth estimation methods:

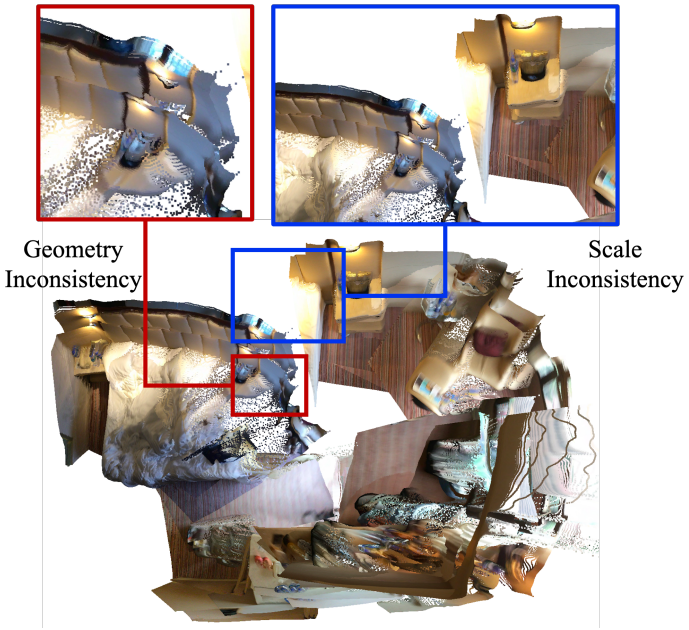
- Geometric inconsistency

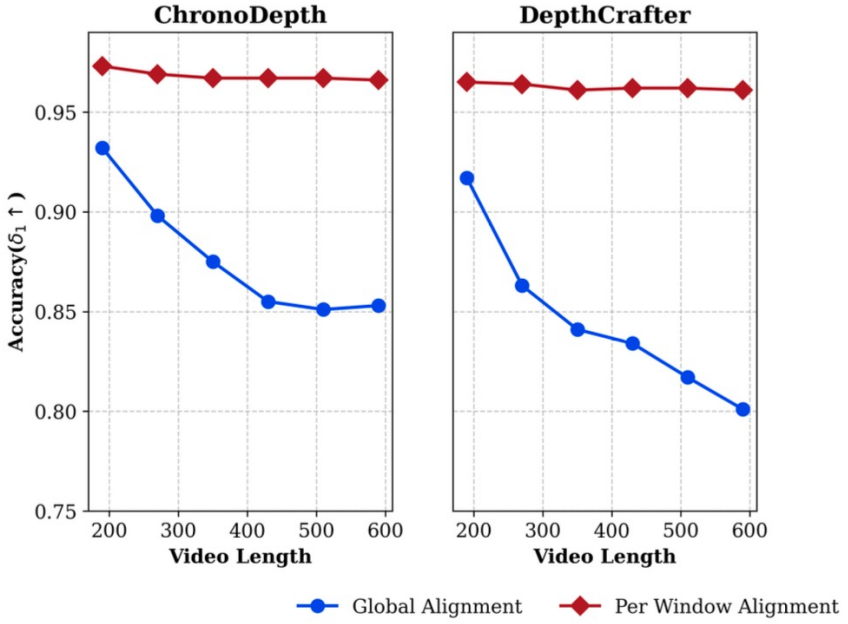
Misalignment across frames when projecting depths to 3D space.

- Scale inconsistency

Sliding window scheme, depth scale varies between windows.







Observation:

- Video depths contain inherent geometry constraints across frames.
- Iterative diffusion denoising process is similar to the traditional iterative geometric optimization process.

Geometry Guidance (Intra Window)

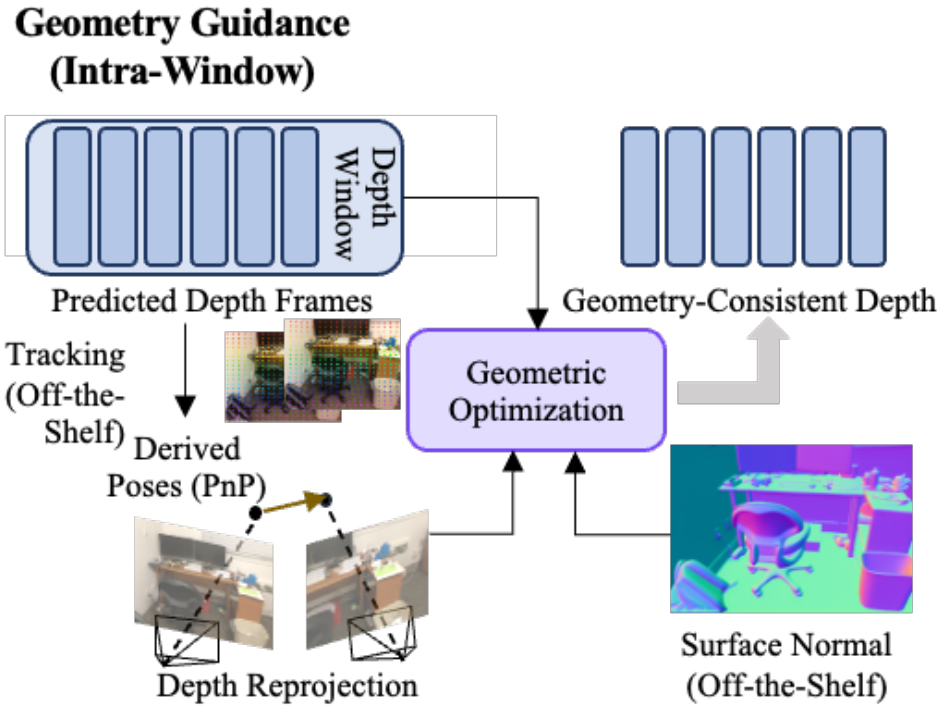
- Input: predicted clean depth at each denoising step

$$\hat{z}_0 = \frac{z_t - (\sqrt{1-\alpha_t})\epsilon_\theta(z_t, t)}{\sqrt{\alpha_t}}$$

- Backward guidance process:

1) $\hat{z}_0 = \frac{z_t - (\sqrt{1-\alpha_t})\epsilon_\theta(z_t, t)}{\sqrt{\alpha_t}}$

2 $\epsilon_{\text{aligned}} = \frac{z_t - \sqrt{\alpha_t} z_{\text{aligned}}}{\sqrt{1-\alpha_t}}$



- Geometry Constraints as $\mathcal{L}$
  video depths + off-the-shelf 2D tracking -> derived poses
  depth reprojection loss + tracking loss + surface normal loss

- Interleaved Diffusion and Geometric Optimization Scheme:
  Denoising -> Geometry optimization … -> Denoising -> Geometry optimization

Scale Guidance (Cross Window)

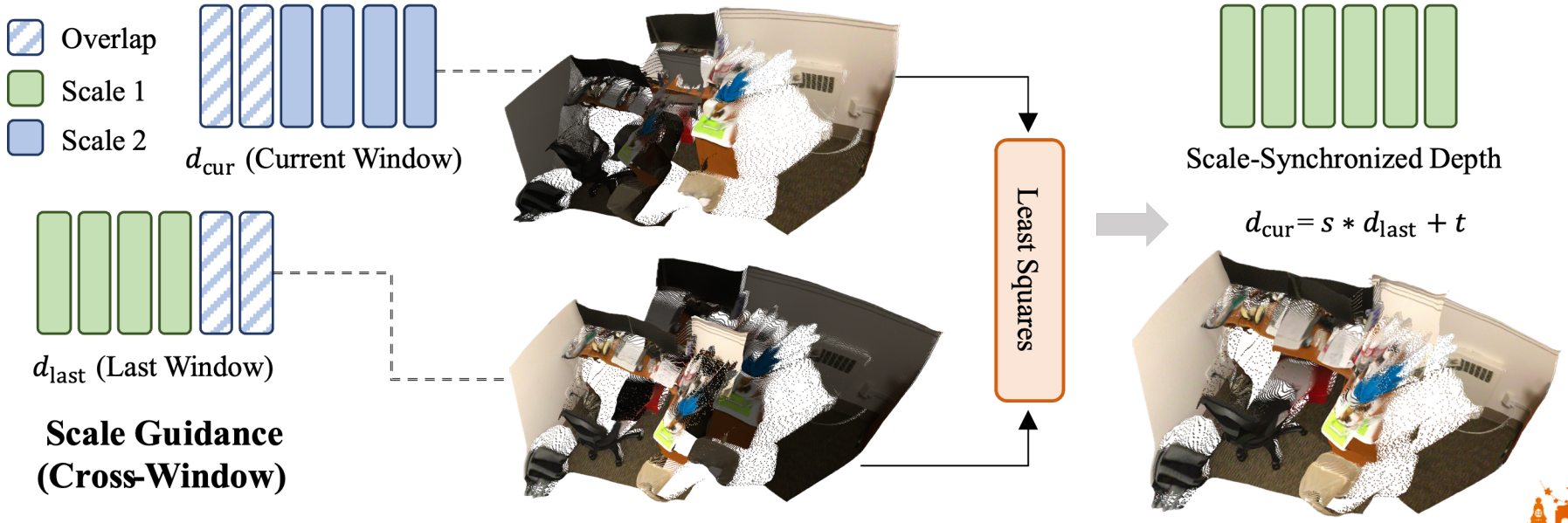
- Existing method: initialize overlap region with last window prediction (**Before** denoising) -> Not Enough !!
- **Key insight**: Impose explicit scale regularization **during** denoising process

- Forward Diffusion Guidance

1) $\hat{\epsilon}_\theta(z_t, t) = \epsilon_\theta(z_t, t) + s(t) \cdot \nabla_{z_t} \mathcal{L}(c, f(\hat{z}_0))$

2) MSE loss between $\mathcal{Z}_{\text{aligned}}$ and $\mathcal{Z}_{\text{cur}}$ as $\mathcal{L}$

- Computation of $\mathcal{Z}_{\text{aligned}}$

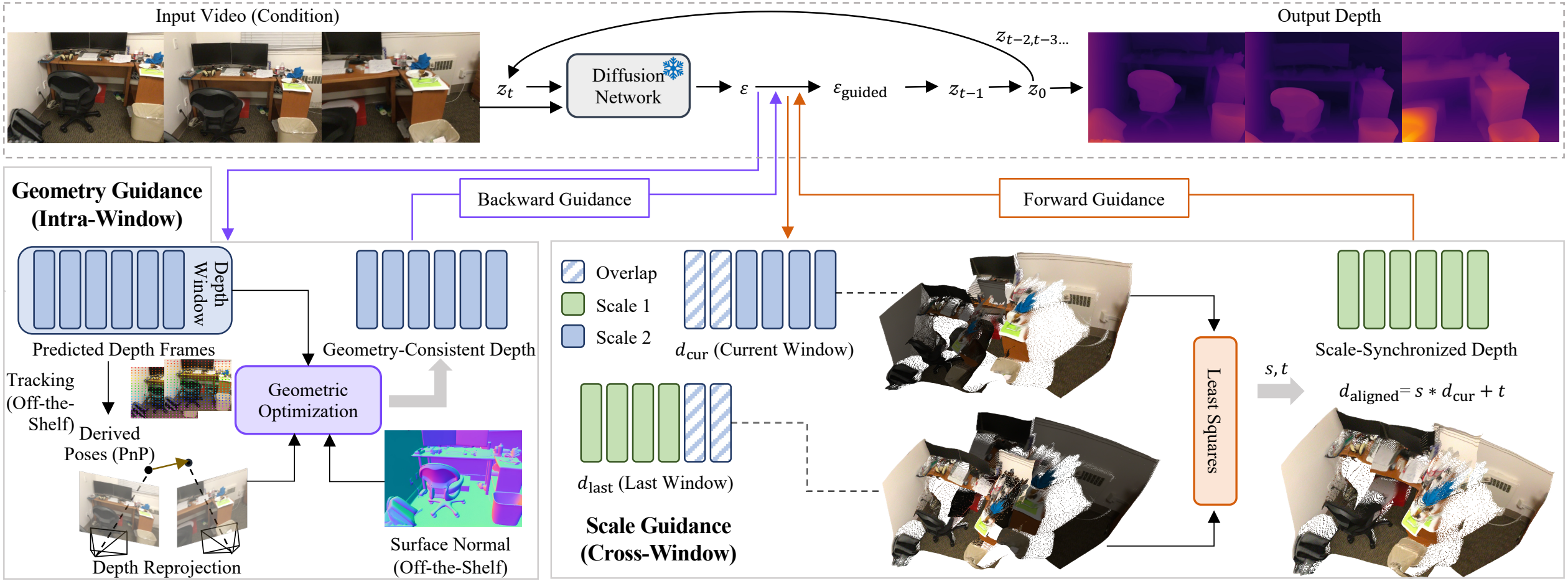


Overall Pipeline

- A **training-free solution** based on Diffusion Posterior Sampling (Diffusion Guidance).

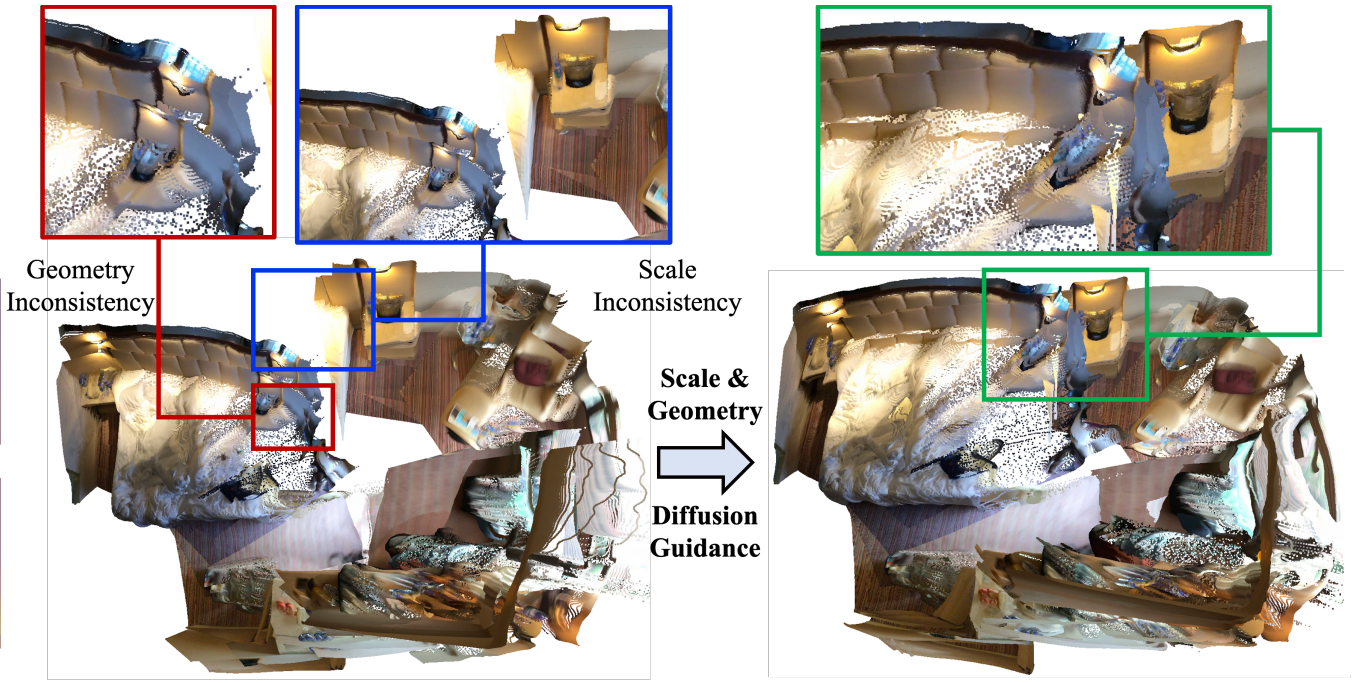
Noise prediction ⇔ score function -$\nabla_x \log p(x)$: $\nabla_x \log p(x|y) = \nabla_x \log p(x) + \nabla_x \log p(y|x)$ ($\approx \nabla_x \log p(y|x_0) = -\nabla_{x_t} L(y, x_0)$)

- A novel optimization framework, which **interleaves** diffusion denoising with conventional geometric and scale optimization for mutual enhancement.

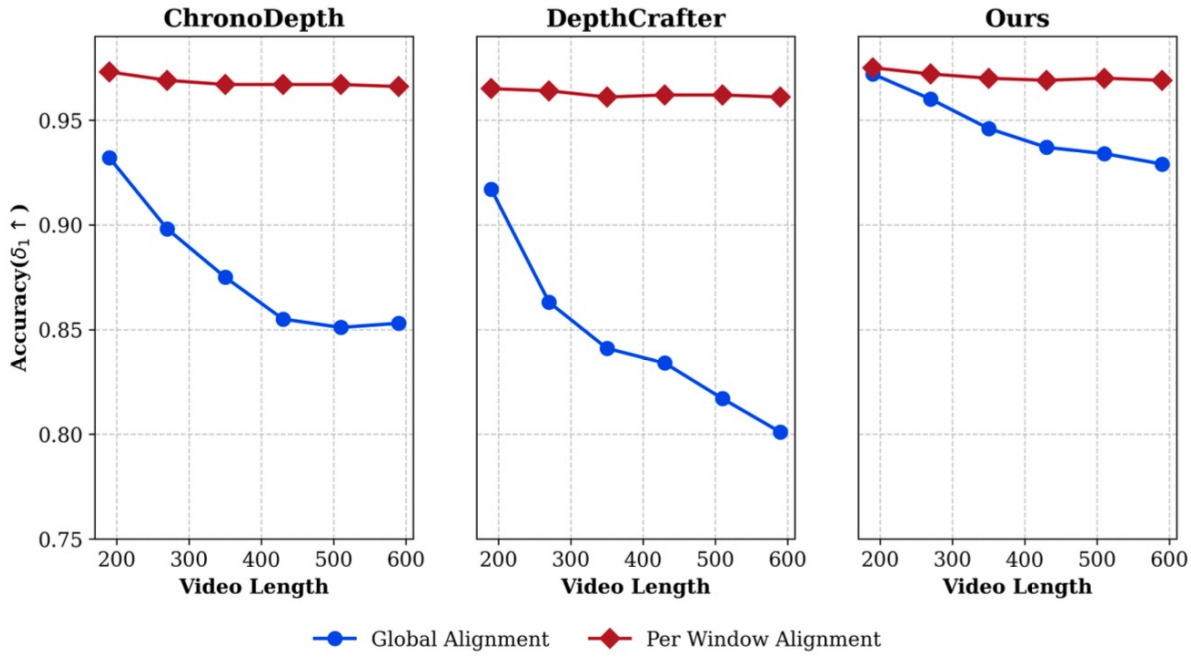


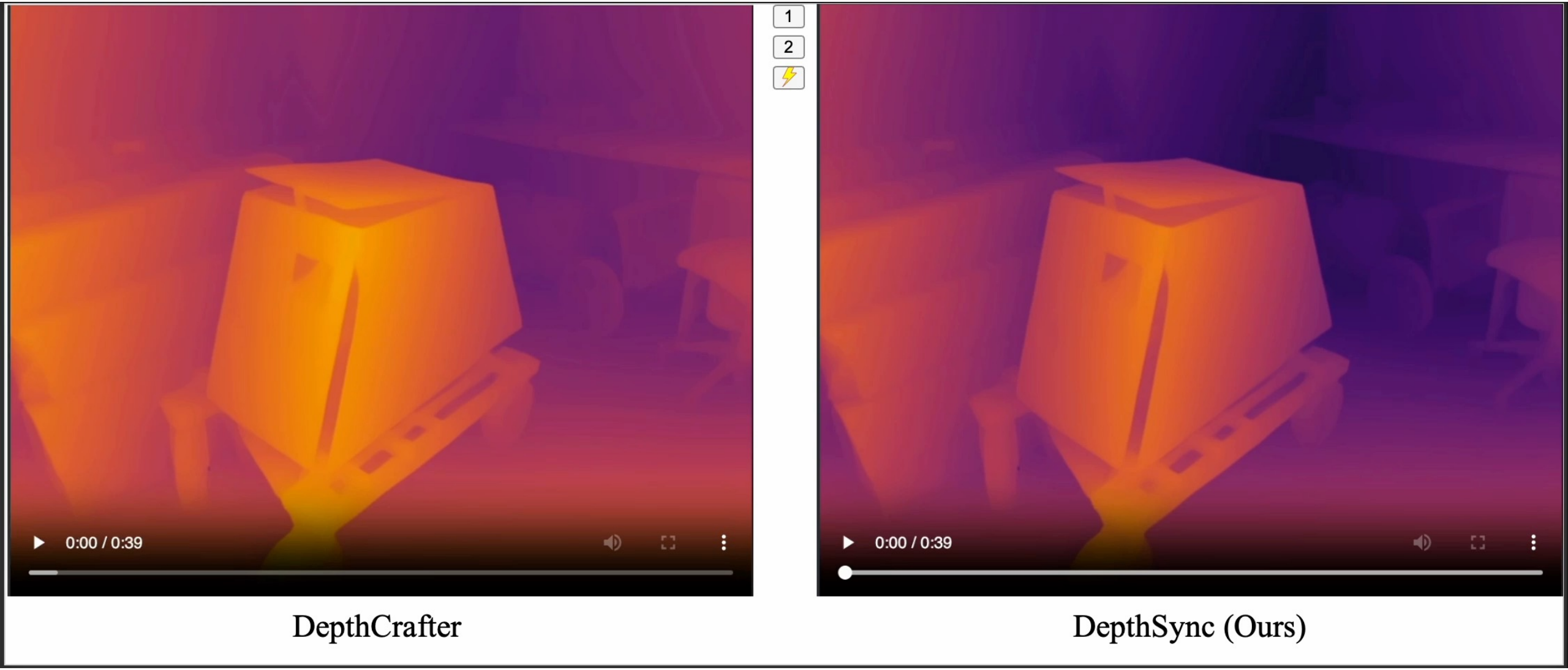
Effectiveness of DepthSync

Geometric Consistency



Scale Consistency





The full video is on the project homepage: https://yuejiangdong.github.io/depthsync/

Long Video Depth Evaluation on Four Benchmarks

Frames per Window			90					110					90
Method	Type	Video Length	ScanNet		GMU KITCHEN		Video Length	KITTI		Bonn		Video Length	ScanNet
			AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑		AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑		MFC↓
DepthAnythingV2	S	150	0.148	0.776	0.188	0.645	190	0.150	0.788	0.102	0.916	150	0.040
Marigold	S		0.179	0.740	0.214	0.584		0.143	0.820	0.098	0.932		0.131
NVDS	V		0.199	0.647	0.231	0.542		0.212	0.658	0.163	0.771		0.047
ChronoDepth	V		0.172	0.749	0.196	0.650		0.178	0.733	0.092	0.932		0.028
DepthCrafter	V		0.141	0.799	0.143	0.795		0.114	0.879	0.095	0.917		0.024
DepthSync(Ours)	V		**0.113**	**0.870**	**0.113**	**0.881**		**0.110**	**0.887**	**0.069**	**0.972**		**0.019**
DepthAnythingV2	S	210	0.160	0.743	0.189	0.652	270	0.156	0.777	0.117	0.887	210	0.040
Marigold	S		0.191	0.710	0.221	0.598		0.146	0.809	0.113	0.901		0.143
NVDS	V		0.208	0.622	0.223	0.578		0.217	0.647	0.176	0.741		0.048
ChronoDepth	V		0.182	0.726	0.201	0.659		0.185	0.707	0.112	0.898		0.028
DepthCrafter	V		0.156	0.757	0.155	0.761		0.116	0.867	0.121	0.863		0.025
DepthSync(Ours)	V		**0.127**	**0.836**	**0.113**	**0.884**		**0.111**	**0.882**	**0.077**	**0.960**		**0.021**
DepthAnythingV2	S	270	0.161	0.738	0.181	0.681	350	0.161	0.760	0.125	0.870	270	0.041
Marigold	S		0.195	0.699	0.219	0.613		0.151	0.796	0.123	0.884		0.148
NVDS	V		0.211	0.614	0.222	0.586		0.219	0.643	0.180	0.722		0.048
ChronoDepth	V		0.184	0.716	0.201	0.659		0.185	0.705	0.124	0.875		0.029
DepthCrafter	V		0.161	0.745	0.154	0.764		0.119	0.859	0.130	0.841		0.025
DepthSync(Ours)	V		**0.136**	**0.814**	**0.120**	**0.860**		**0.112**	**0.874**	**0.083**	**0.946**		**0.021**
DepthAnythingV2	S	330	0.164	0.729	0.191	0.655	430	0.160	0.762	0.127	0.861	330	0.041
Marigold	S		0.198	0.691	0.228	0.611		0.150	0.800	0.126	0.871		0.166
NVDS	V		0.214	0.608	0.229	0.568		0.221	0.641	0.182	0.712		0.049
ChronoDepth	V		0.187	0.709	0.217	0.621		0.186	0.706	0.130	0.855		0.029
DepthCrafter	V		0.168	0.725	0.153	0.756		0.123	0.853	0.134	0.834		0.025
DepthSync(Ours)	V		**0.138**	**0.802**	**0.128**	**0.842**		**0.115**	**0.871**	**0.090**	**0.937**		**0.021**
DepthAnythingV2	S	390	0.167	0.724	0.193	0.645	510	0.158	0.766	0.128	0.861	390	0.041
Marigold	S		0.202	0.689	0.228	0.613		0.149	0.800	0.126	0.871		0.148
NVDS	V		0.216	0.603	0.232	0.561		0.221	0.641	0.185	0.701		0.050
ChronoDepth	V		0.190	0.704	0.222	0.619		0.186	0.705	0.132	0.851		0.030
DepthCrafter	V		0.169	0.721	0.161	0.746		0.124	0.850	0.138	0.817		0.025
DepthSync(Ours)	V		**0.150**	**0.780**	**0.128**	**0.848**		**0.116**	**0.870**	**0.091**	**0.934**		**0.021**
DepthAnythingV2	S	450	0.169	0.722	0.193	0.644	590	0.158	0.769	0.131	0.856	450	0.041
Marigold	S		0.205	0.686	0.241	0.597		0.148	0.803	0.128	0.865		0.147
NVDS	V		0.219	0.597	0.235	0.558		0.223	0.640	0.189	0.692		0.050
ChronoDepth	V		0.193	0.699	0.225	0.612		0.193	0.691	0.132	0.853		0.030
DepthCrafter	V		0.171	0.716	0.160	0.748		0.173	0.727	0.143	0.801		0.025
DepthSync(Ours)	V		**0.154**	**0.769**	**0.131**	**0.837**		**0.117**	**0.869**	**0.093**	**0.929**		**0.021**

Long Video Pose (Derived from Depth) Evaluation on ScanNet

Length	Metric	RelPose++	PoseDiff	RayDiff	DC	**Ours**
150	ATE (m)↓	0.604	0.570	0.522	0.154	**0.144**
	RPE trans (m)↓	0.078	0.132	0.176	0.021	**0.020**
	RPE rot (deg)↓	2.91	3.41	26.9	0.622	**0.611**
210	ATE (m)↓	0.741	0.737	0.666	0.211	**0.197**
	RPE trans (m)↓	0.093	0.149	0.197	0.022	**0.020**
	RPE rot (deg)↓	2.96	4.09	28.3	0.624	**0.620**
270	ATE (m)↓	0.846	0.854	0.792	0.256	**0.242**
	RPE trans (m)↓	0.098	0.156	0.198	0.023	**0.021**
	RPE rot (deg)↓	2.96	4.11	29.4	0.641	**0.635**
330	ATE (m)↓	0.899	0.906	0.849	0.292	**0.275**
	RPE trans (m)↓	0.093	0.158	0.200	0.023	**0.021**
	RPE rot (deg)↓	2.97	4.11	30.3	0.642	**0.627**
390	ATE (m)↓	0.941	0.954	0.891	0.334	**0.316**
	RPE trans (m)↓	0.092	0.152	0.204	0.023	**0.021**
	RPE rot (deg)↓	2.99	4.12	30.5	0.650	**0.643**
450	ATE (m)↓	1.00	1.01	0.942	0.367	**0.350**
	RPE trans (m)↓	0.094	0.162	0.209	0.023	**0.021**
	RPE rot (deg)↓	3.12	4.29	30.7	0.676	**0.669**

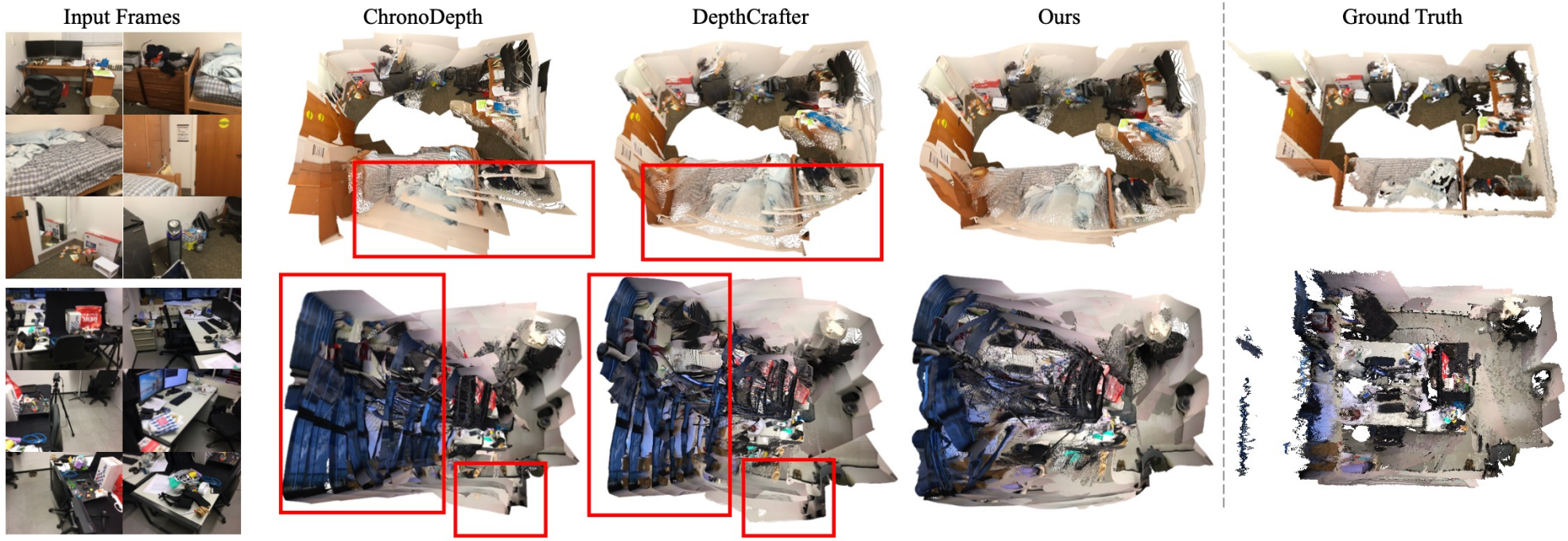
Ablation Study:

Strategy	AbsRel↓	δ_1 ↑	MFC
Baseline	0.174	0.702	0.024
Guidance Term Ablation			
Scale Guidance Only	0.151	0.772	0.024
Geometry Guidance Only	0.166	0.720	0.018
Post Optimization			
Post Scale Alignment	0.157	0.750	0.024
Post Geometry Optimization	0.175	0.694	0.020
Post Scale & Geometry	0.151	0.757	0.020
Ours	**0.137**	**0.801**	**0.018**

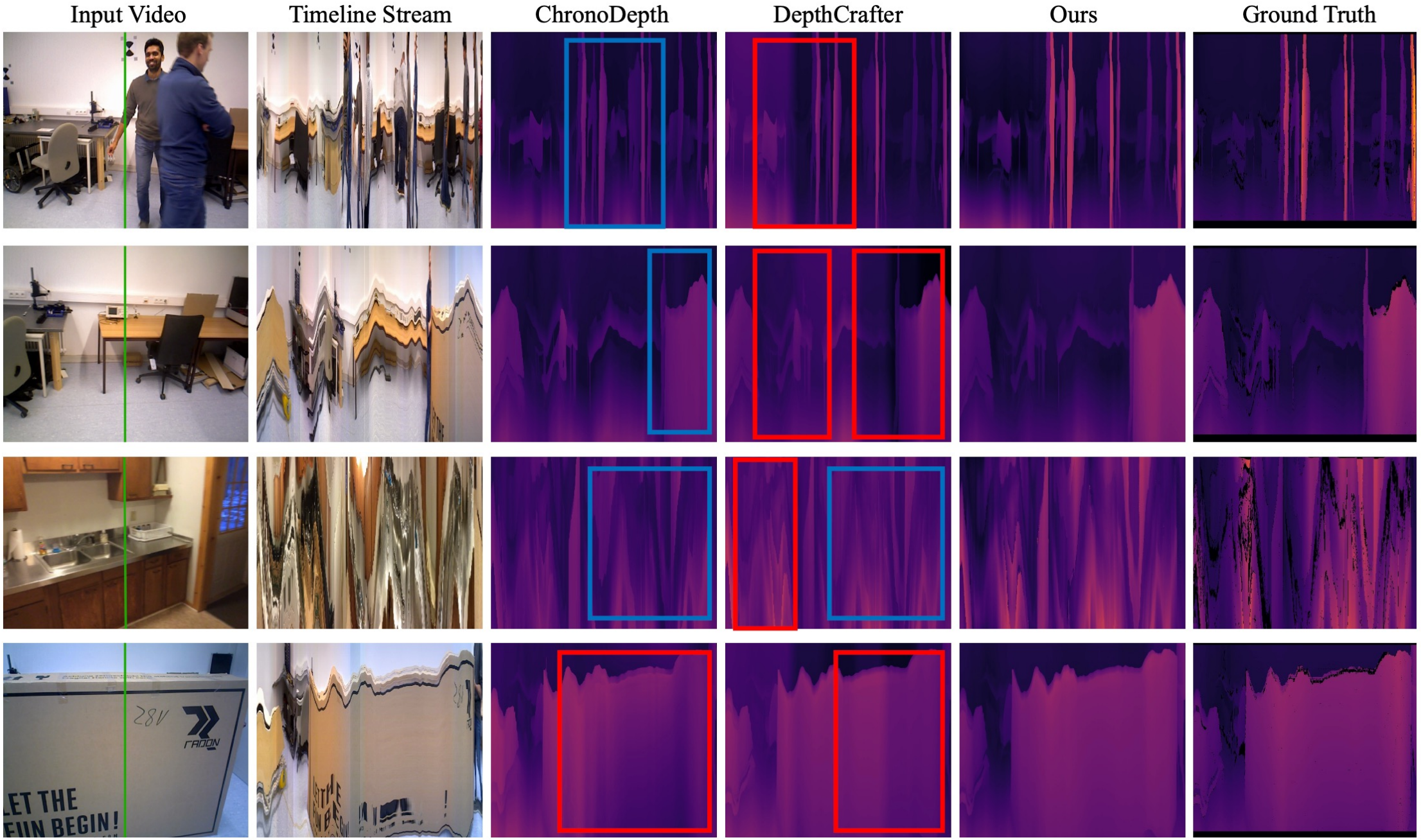
Optimization in the denoising loop is better than pure post-optimization after diffusion.

Geometry Consistency

Qualitative Result



Scale Consistency Qualitative Result



Thank You!

DepthSync: Diffusion Guidance-Based Depth Synchronization for Scale- and Geometry-Consistent Video Depth Estimation

Yue-Jiang Dong[1], Wang Zhao[2], Jiale Xu[2], Ying Shan[2], Song-Hai Zhang[1]

[1]Tsinghua University [2]ARC Lab, Tencent PCG