

Learning Temporally Consistent Video Depth from Video Diffusion Priors

Supplementary Material

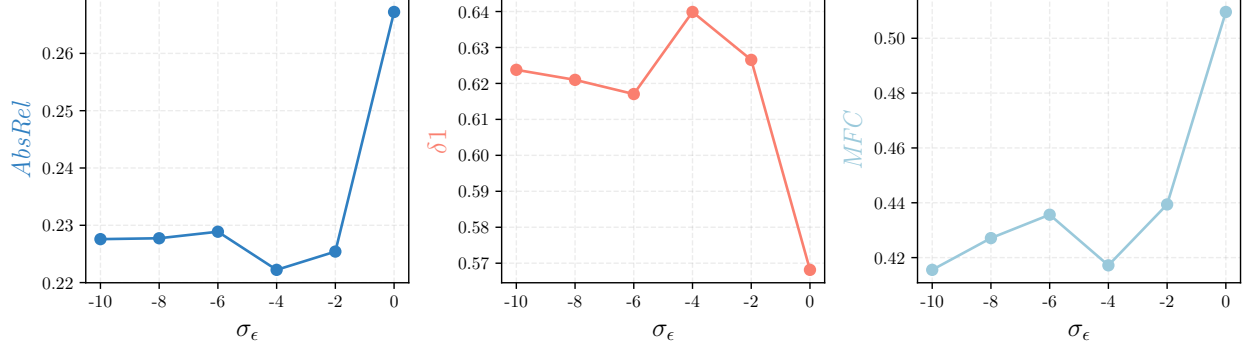


Figure S1. **Ablation Study.** We report accuracy and consistency metrics of our method on *KITTI-360* with different σ_ϵ .

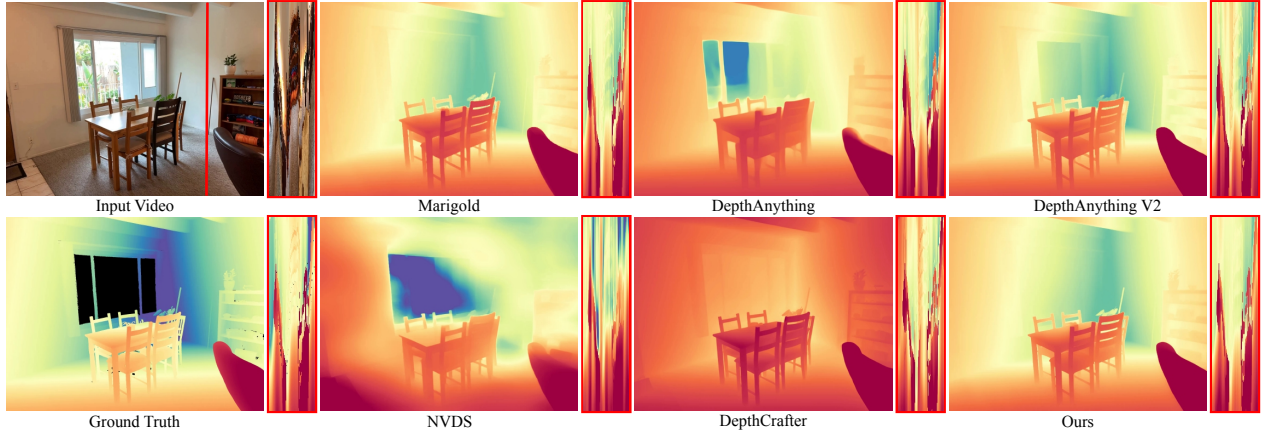


Figure S2. **More qualitative comparison for video depth estimation.** Results on *ScanNet++* dataset.

A. Details on Evaluation Protocol

Datasets. The sequences on *KITTI-360* we chose follows the rules below. Given the absence of ground-truth poses for the initial frames of each sequence, we extracted frames 300-500, ultimately utilizing 200 frames for evaluation.

Metrics. To measure *temporal consistency*, we introduce multi-frame consistency (*MFC*): given two depth maps $D^m, D^n \in \mathcal{R}^{W \times H}$ at frame m, n of the video sequence, we unproject D^m into a point cloud; using the ground-truth world-to-camera poses $P_m, P_n \in \mathcal{R}^{3 \times 4}$ for frames m, n , camera, we transform the point cloud from frame m ' camera space to frame n ' camera space, and project it onto frame n 's image plane to yield $D^{m \rightarrow n}$. We measure temporal consistency as the average L1 distance between $D^{m \rightarrow n}$ and D^n . We mask out invalid pixels in both frames. In practice, we calculate multi-frame consistency on adjacent frames.

B. Detailed Proof on Mathematical Rigor and Fluctuating Guidance

To ensure enough context information, we aim to sample depth latents $\hat{\mathbf{z}}^{(\mathbf{d}_{W:F})}$ conditioned on $\hat{\mathbf{z}}^{(\mathbf{d}_{0:W})}$, which is $p_\theta(\hat{\mathbf{z}}^{(\mathbf{d}_{W:F})} | \hat{\mathbf{z}}^{(\mathbf{d}_{0:W})})$. **For the replacement trick**, the sampling of $\hat{\mathbf{z}}^{(\mathbf{d}_{W:F})}$ follows standard unconditional sampling from $p_\theta(\hat{\mathbf{z}}^{(\mathbf{d}_{0:F})} | \hat{\mathbf{z}}^{(\mathbf{d}_{0:F})})$, where $\hat{\mathbf{z}}^{(\mathbf{d}_{0:F})} = [\hat{\mathbf{z}}^{(\mathbf{d}_{0:W})}, \hat{\mathbf{z}}^{(\mathbf{d}_{W:F})}]$. Crucially, samples $\hat{\mathbf{z}}^{(\mathbf{d}_{0:W})}$ are replaced at each step by exact forward process samples $q(\hat{\mathbf{z}}^{(\mathbf{d}_{0:W})} | \hat{\mathbf{z}}^{(\mathbf{d}_{0:W})})$. This causes to update $\hat{\mathbf{z}}^{(\mathbf{d}_{W:F})}$ using $\hat{\mathbf{z}}^{(\mathbf{d}_{W:F})}(\hat{\mathbf{z}}^{(\mathbf{d}_{W:F})}) \approx \mathbb{E}_q[\hat{\mathbf{z}}^{(\mathbf{d}_{W:F})} | \hat{\mathbf{z}}^{(\mathbf{d}_{W:F})}, \hat{\mathbf{z}}^{(\mathbf{d}_{0:W})}]$, while what is needed instead is $\mathbb{E}_q[\hat{\mathbf{z}}^{(\mathbf{d}_{W:F})} | \hat{\mathbf{z}}^{(\mathbf{d}_{W:F})}, \hat{\mathbf{z}}^{(\mathbf{d}_{0:W})}] = \mathbb{E}_q[\hat{\mathbf{z}}^{(\mathbf{d}_{W:F})} | \hat{\mathbf{z}}^{(\mathbf{d}_{0:F})}, \hat{\mathbf{z}}^{(\mathbf{d}_{0:W})}] + (\sigma_t^2 / \alpha_t) \nabla_{\hat{\mathbf{z}}^{(\mathbf{d}_{W:F})}} \log q(\hat{\mathbf{z}}^{(\mathbf{d}_{0:W})} | \hat{\mathbf{z}}^{(\mathbf{d}_{0:W})})$. The missing second term introduces dynamic guidance variations across sampling steps. **As for our context-aware strategy**, we

		Marigold	DAv2	NVDS	DC	Ours
Inference Speed (s)		5.64	0.80	1.05	1.30	0.49
Compute (GB)		5.67	23.7	20.5	8.04	6.6
# of Parameters (B)		1.29	0.33	0.35	2.25	2.25
Training data	# of frames	74K	62.6M	1.4M	-	39K
	# of scenes	-	-	14.2K	203K	938

Table S1. Speed and compute comparison.

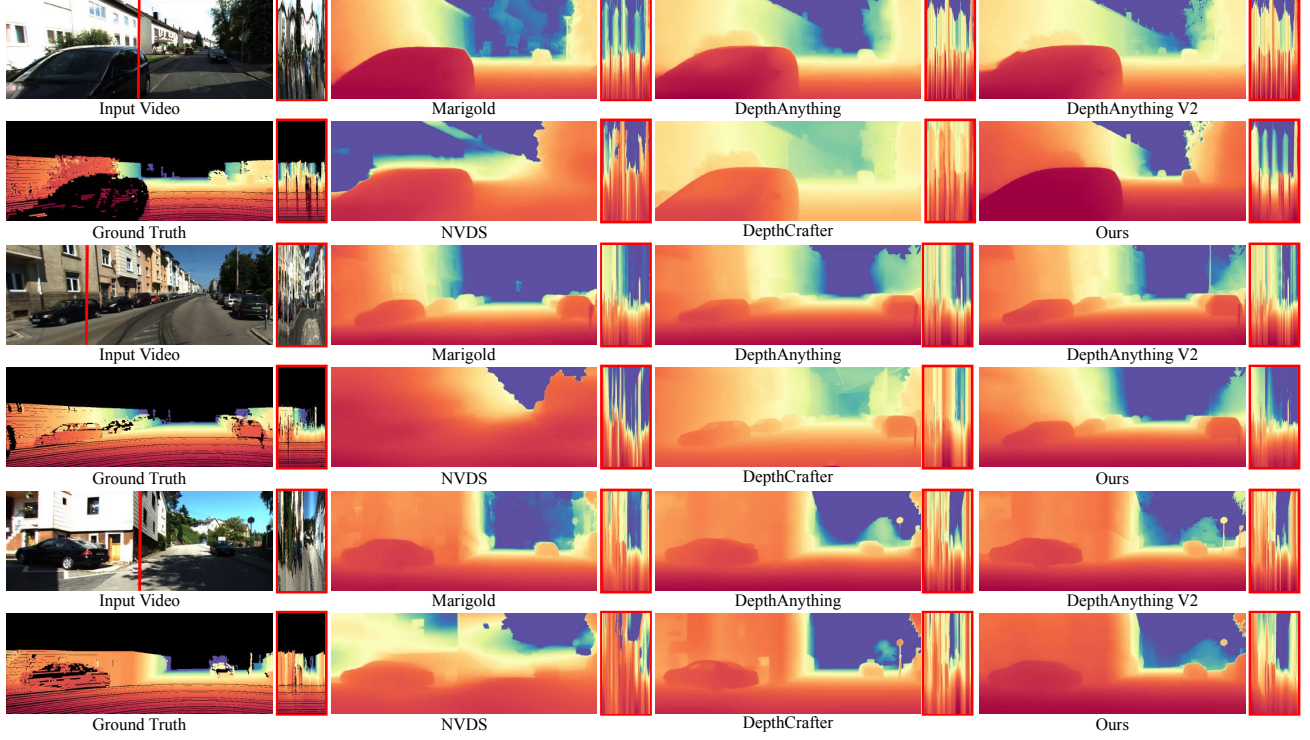


Figure S3. More qualitative comparison for video depth estimation. Results on KITTI-360 datasets.

can do conditional sampling from $p_\theta(\hat{\mathbf{z}}_{t-1}^{(\mathbf{d}_{0:F})} | \hat{\mathbf{z}}_t^{(\mathbf{d}_{0:F})})$, with $\hat{\mathbf{z}}_t^{(\mathbf{d}_{0:F})} = [\hat{\mathbf{z}}_t^{(\mathbf{d}_{0:W})}, \hat{\mathbf{z}}_t^{(\mathbf{d}_{W:F})}]$ without forward process $q(\cdot|\cdot)$. As a result, $\hat{\mathbf{z}}_{t-1}^{(\mathbf{d}_{0:F})}$ is updated in the direction provided by $\mathbb{E}[\hat{\mathbf{z}}_{t-1}^{(\mathbf{d}_{W:F})} | \hat{\mathbf{z}}_t^{(\mathbf{d}_{W:F})}, \hat{\mathbf{z}}_t^{(\mathbf{d}_{0:W})}]$.

C. Speed and Compute Comparison

Tab. S1 shows runtime, compute and model parameters. ChronoDepth is significantly faster than Marigold [40] and DepthCrafter (DC) [33], and requires a fraction of the memory used by DepthAnything v2(DAv2) [76], thanks to our more lightweight UNet architecture compared with the baselines.

D. Additional Ablation

We investigate the significance of the small noise level σ_ϵ in the context of overlapping frames within arbitrarily long videos. As illustrated in Fig. S1, an excessively small σ_ϵ

results in degraded spatial and temporal performance due to compounded errors. Conversely, an overly large σ_ϵ also leads to diminished spatial and temporal performance. Consequently, we opt for $\sigma_\epsilon = -4$.

E. More Qualitative Results

We provide more qualitative comparisons from KITTI-360, ScanNet++ and Bonn datasets in Figs. S2 to S4. First, we highlight the remarkable spatial accuracy achieved by our method, being comparable to or even better than the one by state-of-the-art models. Furthermore, we can notice how the y-t slice by most methods shows high-frequency artifacts, whereas ours is consistently smoother, confirming the superior temporal consistency we achieve.

F. Limitation

Our method is robust to rapid ego-camera motion (ScanNet++) and long video (KITTI-360). However, we observe



Figure S4. **More qualitative comparison for video depth estimation.** Results on *ScanNet++* and *Bonn* datasets.

a slight degradation in *AbsRel* when handling scenes with abundant dynamic objects (*Sintel*). We attribute this to the limited moving objects in the training data, which can be extended in the future.