ICCV LongVid Foundations Workshop #5

StereoDiff: Stereo-Diffusion Synergy for Video Depth Estimation

Anonymous ICCV LongVid Foundations Workshop submission

Paper ID 5

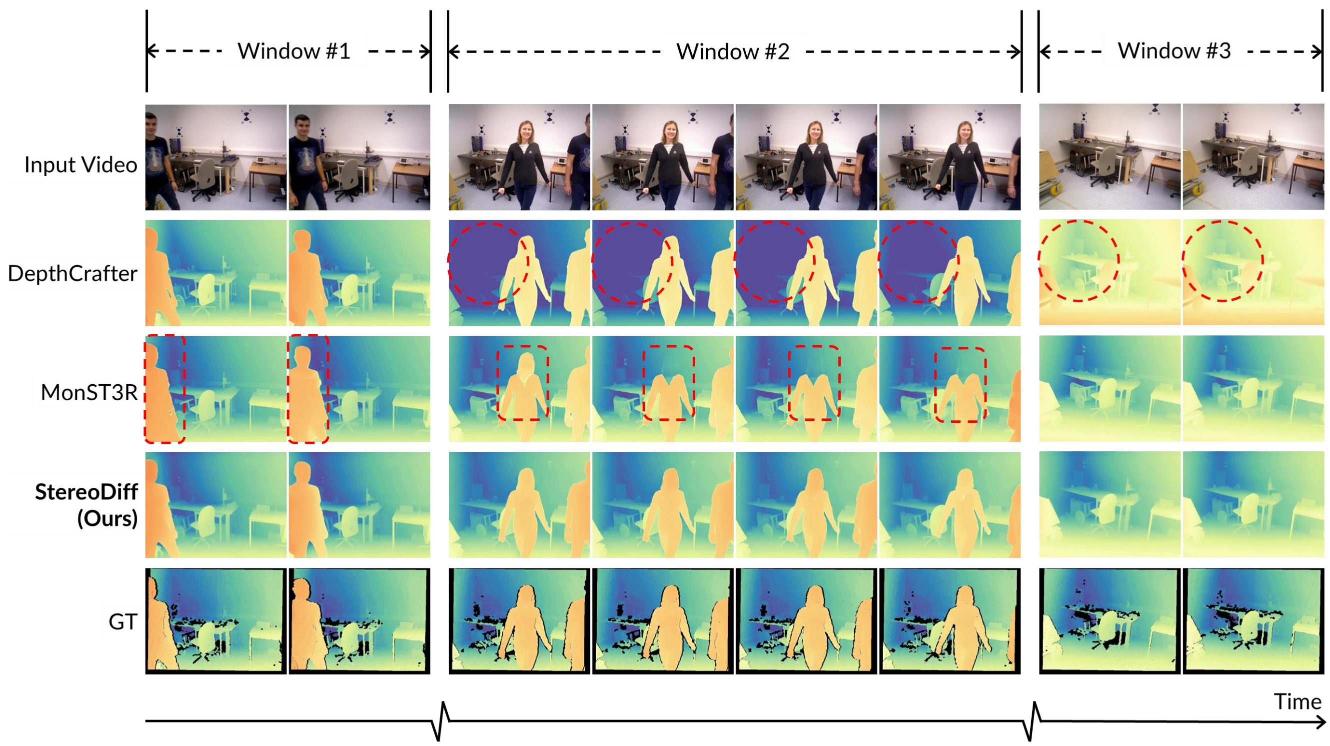


Figure 1. **StereoDiff excels in delivering remarkable global and local consistency for video depth estimation.** In terms of global consistency, StereoDiff achieves highly accurate and stable depth maps on static backgrounds across consecutive windows, leveraging stereo matching to prevent the abrupt depth shifts often seen in DepthCrafter [32], where depth values on static backgrounds can vary significantly between adjacent windows. For local consistency, StereoDiff yields much smoother, flicker-free depth values across consecutive frames, especially in dynamic regions. In contrast, MonST3R [91] suffers from frequent, pronounced flickering and jitters in these areas.

Abstract

Recent video depth estimation methods achieve great performance by following the paradigm of image depth estimation, i.e., typically fine-tuning pre-trained video diffusion models with massive data. However, we argue that video depth estimation is not a naive extension of image depth estimation. The temporal consistency requirements for dynamic and static regions in videos are fundamentally different. Consistent video depth in static regions, typically backgrounds, can be more effectively achieved via stereo matching across all frames, which provides much stronger global 3D cues. While the consistency for dynamic regions still should be learned from large-scale video depth data to ensure smooth transitions, due to the violation of triangulation. Based on these insights, we introduce ***StereoDiff****, a two-stage video depth estimator that synergizes stereo matching for mainly the static areas with video depth diffusion for maintaining consistent depth transitions in dynamic areas. We mathematically demonstrate how stereo matching and video depth diffusion offer complementary strengths through frequency domain analysis, highlighting the effectiveness of their synergy in capturing the advantages of both. Experimental results on zero-shot, real-world, dynamic video depth bench-*

marks, both indoor and outdoor, demonstrate StereoDiff's SoTA performance, showcasing its superior consistency and accuracy in video depth estimation. Project page: link.

1. Introduction

Monocular video depth estimation is a foundational task in 3D computer vision. Particularly after the hot trend of leveraging pre-trained Stable Diffusion (SD) [54] for image depth prediction [19, 22, 27, 33, 44], *e.g.*, Marigold [33] and Lotus [27], we have witnessed emerging attentions on video depth estimation in the community [17, 32, 36, 61, 72, 83, 91]. Many of them fine-tune the Stable Video Diffusion (SVD) [4] using large-scale video depth data, *e.g.*, DepthCrafter [32] and DepthAnyVideo [17]. However, most previous methods [17, 32, 61, 75, 82, 83] consider the video depth estimation merely as a video version of image depth estimator, directly modeling a mapping function from the RGB video distribution to the video depth distribution, similar to previous image depth methods that fit a mapping function directly from image distribution to depth.

In this paper, we argue that *video depth estimator is not simply a video version of image depth estimator*. The core attribute of video depth estimation is *consistency*. The consistency for dynamic and static parts of the scene is essentially different and should be handled separately.

① Static regions involve only the camera motion, allowing the 3D structure to be analytically inferred from pairwise correspondences obtained through stereo matching [23, 46, 58, 59, 71, 72, 76, 91] on a sequence of RGB frames, providing strong global 3D cues. The consistency of these areas, primarily about static backgrounds and across all video frames, is termed *global consistency*. Since static elements often occupy a large portion of the scene (*e.g.*, roads, trees, buildings outdoors, or walls, tables, and floors indoors), a strong and robust global consistency is the foundation for achieving consistent and accurate video depth estimation. ② Dynamic parts contain both object motions and camera motion. It is infeasible to achieve analytical 4D reconstruction from RGB sequence alone, as it requires solving unknown object shapes, poses, and motion trajectories simultaneously, which is highly ill-posed. For example, imagine a scene where a person is waving his/her hand from left to right. The predicted depth maps are expected to not only strictly correspond to the RGB inputs in image composition, but also more importantly, maintain consistent, smooth depth changes for the moving hand across consecutive frames, without abrupt fluctuations or flickering. This temporal consistency across short sequences and particularly in dynamic areas, is termed *local consistency*, which should be learned by seeing large amount of video depth data.

Motivated by these analysis, we propose *StereoDiff*, a novel two-stage video depth estimator that synergizes both the stereo matching [36, 72, 91] for accurate global consistency and a video depth diffusion model [17, 32, 61] fine-tuned on large-scale video depth datasets for smooth local consistency. In the first stage of StereoDiff (Sec. 3.2), all video frames are processed in pairs through a stereo matching pipeline and then merged to establish strong global consistency. However, for dynamic objects, depth predictions are limited to pairwise frames (equivalent to a window size of 2), leading to clear inconsistencies (Fig. 1, middle column). Potential camera motion errors can also cause depth jitters across consecutive frames, resulting in suboptimal local consistency. To tackle this issue, in the second stage of StereoDiff (Sec. 3.3), a one-step video depth diffusion process is employed, in order to greatly improve the local consistency of stereo matching-based depth maps while preserving their original strong global consistency, resulting in video depth maps with both high-quality global and local consistency. Leveraging the priors of pre-trained video diffusion models, *e.g.*, SVD, and fine-tuning them with extensive video depth data, video depth diffusion models achieve exceptionally smooth local consistency across neighboring frames. However, it is typically impossible for video diffusion-based video depth estimators to process all video frames simultaneously, which inherently limits their global consistency, as illustrated in the second column of Fig. 1.

We validate StereoDiff on four *zero-shot* video depth benchmarks (Tab. 1): Bonn [47] (real, dynamic, indoor); KITTI [24] (real, dynamic, outdoor); ScanNetV2 [13] (real, static, indoor); and Sintel [8] (synthetic, dynamic, various). The StereoDiff achieves the *best* comprehensive results. We also report the performance on different frequency domains (Tab. 2) and the performance on static and dynamic regions (Tab. 3), to assess on global and local consistency, respectively. StereoDiff effectively retains the strong global consistency established in the first stage while significantly enhancing the local consistency in the second. Additionally, thanks to the one-step denoising policy in the second stage, StereoDiff is ~ 2.1 times faster than DepthCrafter (Please see Tab. 3 in the Supplementary Materials). The summarized key contributions are:

- We emphasize that achieving consistent video depth estimation requires distinct treatment for static (background) and dynamic (foreground) regions. Specifically, global consistency is better achieved through stereo matching on static regions, while local consistency for dynamic objects should be learned from large-scale video depth data.
- Based on these insights, we introduce *StereoDiff*, a novel two-stage video depth estimator that synergizes stereo matching for strong global consistency and video depth diffusion for smooth local consistency, delivering reliable video depth estimations. StereoDiff is training-free and does not require test-time optimization.
- Experimental results on dynamic, zero-shot, real-world

ICCV LongVid Foundations Workshop #5

video depth benchmarks (Tab. 1), both indoor and outdoor, demonstrate StereoDiff's SoTA performance. In addition, analysis across frequency domains (Fig. 3 and Tab. 2) and in dynamic and static regions (Tab. 3) further shows that StereoDiff effectively integrates the strengths of both stereo matching and video depth diffusion models.

2. Related Works

2.1. Image Depth Estimation

Monocular image depth estimation has advanced significantly from early CNN-based approaches [15, 21, 35, 52, 84, 88] to vision transformer-based [14, 53, 85, 89]. To build powerful and generalizable depth estimators, DepthAnything [80, 81] and Metric3D [31, 86] series leveraged extensive training data comprising millions of samples, achieving SoTA performance. Additionally, some methods [1, 5, 48], *e.g.*, DepthPro [5] focus on accurately estimating the metric depth. Recent SD-based depth predictor, *e.g.*, Marigold [33] and GeoWizard [22] incorporated pre-trained diffusion priors for monocular depth estimation, achieved remarkable zero-shot generalizability. More recent studies [27, 44], *e.g.*, Lotus [27], E2E-FT [44], have further shown that single-step diffusion delivers even superior performance.

2.2. Video Depth Estimation

SfM for Video Depth. Traditional Structure-from-Motion (SfM) methods [18, 20, 58, 59, 62, 69, 77, 95] can estimate only static 3D structure and camera positions, as dynamic objects violate triangulation constraints. Neither can those real-time visual SLAM systems [16, 56, 57, 65, 68], *e.g.*, NeuralRecon [65] and DoubleTake [57]. Earlier approaches [23, 46] adapted SfM for motions with strong assumptions, *e.g.*, rigidity. Recently, self-supervised methods [2, 3, 9, 12, 34, 41, 42, 66, 87, 90, 93] have tackled this via jointly estimating of video depth, camera poses, and motion residuals, *e.g.*, GeoNet [87], CasualSAM [93], and Robust-CVD [34, 42]. However, these methods require resource-intensive test-time optimization (or fine-tuning). More recent advancements, *e.g.*, DUSt3R [72], MASt3R [36], and MonST3R [91], deliver more accurate and robust SfM results given monocular videos in an inference-based manner, even with large motions [91]. All video frames are pairwise processed and then merged, which brings global consistency. Nonetheless, due to their pairwise input mechanism, jitters and flickering between consecutive frames still persist, particularly on dynamic objects.

End-to-end Video Depth Estimators. The performance of traditional end-to-end methods [25, 37–39, 67, 70, 73, 75, 82, 83, 90], *e.g.*, DeepV2D [67], NVDS [75], and FutureDepth [83], are inevitable constrained due to limited training data and model capacity. Recently, benefiting from web-scale image datasets [60], diffusion models [11, 26, 28, 45, 50, 51, 54, 55, 63, 64, 92] have achieved exceptional image generation capability, leading to significant progress in video generation [4, 7, 10, 29, 30, 74, 79, 94], *e.g.*, SVD [4] and Sora [7]. More recently, following the advancements of image depth estimation [19, 22, 27, 33, 44], fine-tuning pre-trained video diffusion models using large-scale video depth data has gained traction [17, 32, 61], *e.g.*, DepthAnyVideo [17] and DepthCrafter [32], producing exceptionally smooth video depth predictions. However, input videos are typically divided into windows (of continuous or interpolated frames) and processed sequentially, which can lead to cross-window consistences due to the absence of global 3D constraints.

Motivated by these methods, StereoDiff synergizes the strengths of both SfM and end-to-end video depth diffusion models, aiming to deliver video depth estimations with both strong global consistency and smooth local consistency.

[1]For clearer visualization, we filtered out low-confidence 3D points from the full point cloud, like those representing the moving yellow balloon.

3. Method

Given a monocular video with a sequence of RGB images $\mathcal{I} = \{I_t\}_{t=0}^{T-1}$, the goal of StereoDiff is to predict consistent depth maps across all video frames. As shown in Fig. 2, StereoDiff is a two-stage video depth estimator designed to achieve both global and local consistency. In the first stage, stereo matching [36, 72, 91] is applied across all frames to establish strong global consistency, *i.e.*, $\mathcal{D}_{\text{s}} = \{D_t^{\text{s}}\}_{t=0}^{T-1} = \Theta_{\text{s}}(\mathcal{I})$. In the second stage, we use a video depth diffusion model [17, 32, 61] to enhance local consistency, particularly for dynamic objects, while preserving the global coherence achieved in the first stage, *i.e.*, $\mathcal{D}_{\text{sd}} = \{D_t^{\text{sd}}\}_{t=0}^{T-1} = \Theta_{\text{d}}(\mathcal{D}_{\text{s}}, \mathcal{I})$. This two-stage approach enables StereoDiff to deliver high-quality video depth that maintain coherence across both static and dynamic regions throughout the video. In Sec. 3.1, we formalize global and local consistency from the perspective of frequency domain analysis. Subsequently, Sec. 3.2 and Sec. 3.3 provide detailed descriptions of each stage.

3.1. Formulation of Consistency

Given a video depth estimation $\hat{\mathcal{D}} = \{\hat{D}_t\}_{t=0}^{T-1}$ and the corresponding GT depth $\mathcal{D}^*$, along with a metric function $f_\epsilon(\cdot)$ to measure the errors between them, we can calculate the sequence of error values:

$$\mathcal{E} = \{\epsilon_t\}_{t=0}^{T-1} = f_\epsilon\left(\mathcal{D}^*, \hat{\mathcal{D}}\right) \tag{1}$$

This error sequence can be represented as a sum of orthogonal waves with different frequencies. In this paper, we use fast Fourier transform (FFT) to compute the Discrete Fourier Transform (DFT) of error sequence $\mathcal{E}$, decomposing it into

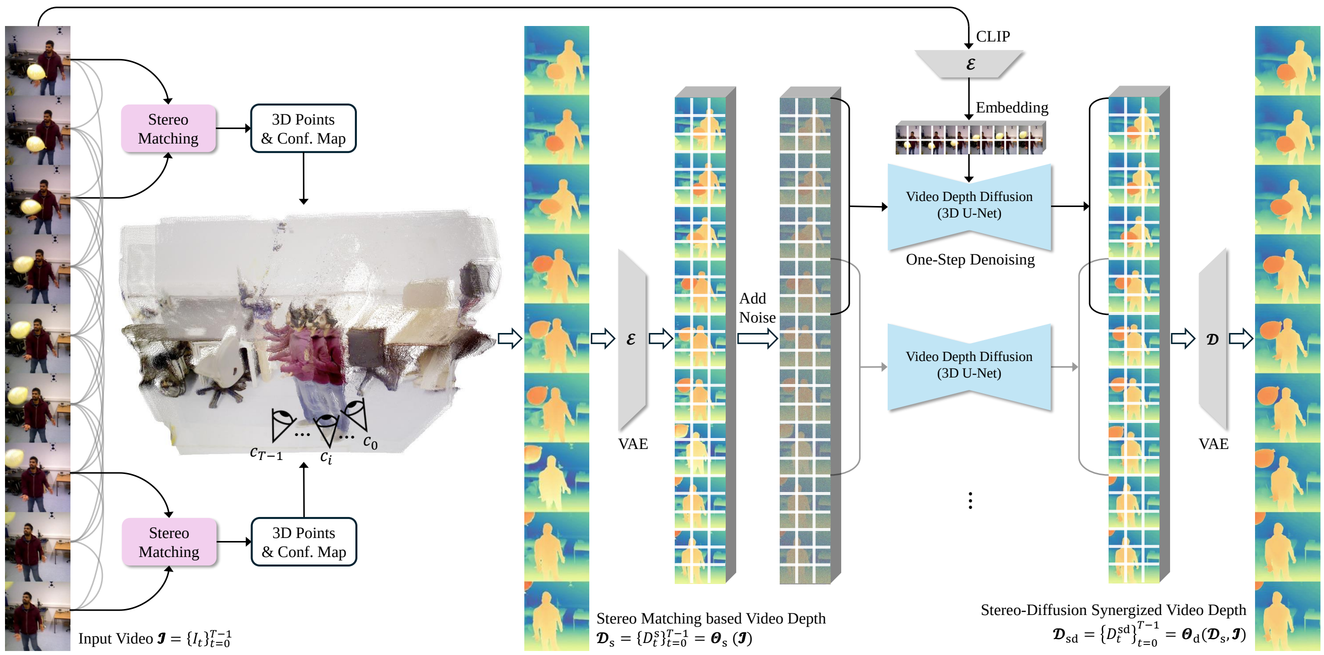


Figure 2. **Pipeline of StereoDiff.** ① All video frames are paired for stereo matching in the first stage, primarily focusing on static backgrounds, in order to achieve a strong global consistency[1]. ② Using the stereo matching-based video depth from the first stage, the second stage of StereoDiff applies a video depth diffusion for significantly improving the local consistency without sacrificing its original global consistency, resulting in video depth estimations with both strong global consistency and smooth local consistency.

several frequency components:

$$\mathcal{F}(\epsilon_k) = \sum_{t=0}^{T-1} \epsilon_t \cdot e^{-i2\pi \frac{k}{T} t}, \quad k = 0, 1, \ldots, T-1 \qquad (2)$$

where $\mathcal{F}(\epsilon_k)$ represents the frequency component at the k-th frequency domain; T is the total number of frames; and i is the imaginary unit. The error sequence can further be reconstructed by Inverse DFT:

$$\epsilon_t = \frac{1}{T} \sum_{k=0}^{T-1} \mathcal{F}(\epsilon_k) \cdot e^{i2\pi \frac{k}{T} t}, \quad t = 0, 1, \ldots, T-1 \qquad (3)$$

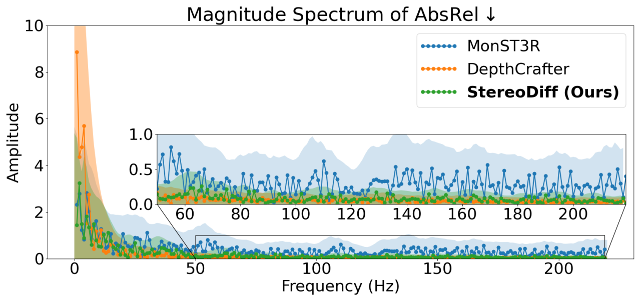
Applying FFT to the error sequence, we can efficiently compute $\mathcal{F}(\epsilon_k)$ for all k frequency domains. This decomposition allows us to analyze the contribution of different frequency bands to the overall error, distinguishing between low-frequency and high-frequency components.

Global consistency refers to the overall stability of depth predictions across the entire video, especially in static backgrounds. For static or minimally dynamic objects, depth changes over time are primarily due to camera motion. Most real-world videos typically have a frame rate much higher than 1 FPS ($\ll$ 1Hz), causing these depth variations to exhibit very low-frequency characteristics, sometimes appearing nearly linear. Global inconsistency often refers to persistent, significant depth deviations that remain stable over long sequences of consecutive frames, which strongly affects the low-frequency components of error sequence $\mathcal{E}$.

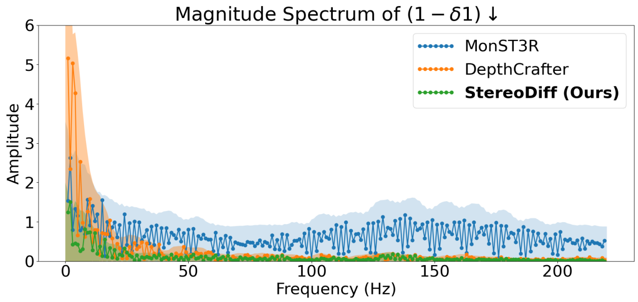
Local consistency focuses on stability between neighboring frames, particularly in dynamic areas with significant motion. Depth variations in these regions are influenced by both camera motion and object motion. Local inconsistencies can arise from: 1) errors in camera motion estimation (common in stereo matching-based methods), causing sudden shifts and depth fluctuations in certain frames; and 2) limited window size, which inevitably prevents consistent and accurate depth tracking of moving objects, resulting in jitters and flickering. Although these local inconsistencies may not be clearly reflected on the overall metrics due to the limited number of affected frames, they can significantly increase the high-frequency amplitudes of the error sequence $\mathcal{E}$.

3.2. Stereo Matching for Global Consistency

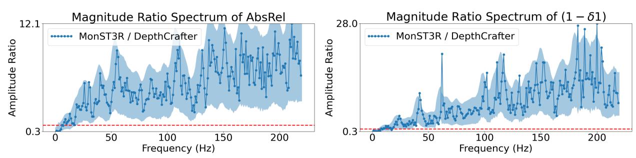
Given the input RGB frames $\mathcal{I}$, the first stage of StereoDiff pairs each frame with the subsequent n frames, forming a total of $nT - (1 + 2 + \cdots + n) = nT - (n+1)n/2$ image pairs. Each pair is then processed through a stereo matching pipeline, resulting in coarse 3D point clouds that ensure the strong global consistency in video depth estimation. Thanks to the advances of SfM [23, 46, 58, 59, 62, 71, 72, 76, 77, 91], we are fortunate to have works like DUSt3R [72], MASt3R [36], and MonST3R [91] that offer highly accurate and robust stereo matching correspondences



(a) Absolute mean relative error (AbsRel) ↓



(b) Inverse accuracy metric $(1-\delta 1)$ ↓ to align with FFT settings.



(c) Amplitude ratio (MonST3R / DepthCrafter) on AbsRel and $(1-\delta 1)$, with the red dotted line denoting the amplitude ratio $= 1$.

Figure 3. **Magnitude spectrum of the error sequence on Bonn [47] dataset.** The first scene of Bonn, "balloon", containing 438 frames, is used as an example here. Due to symmetry, only the second half of the frequency spectrum is shown.

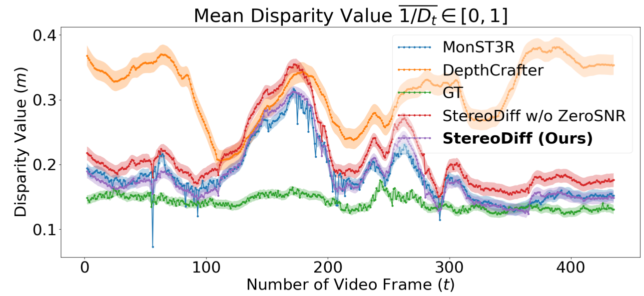


Figure 4. **Comparison of mean disparity[5] value $\overline{1/D_t}$ tested on Bonn [47] dataset** for MonST3R [91], DepthCrafter [32], and StereoDiff. All disparity maps are normalized to $[0, 1]$ on a per-scene basis before comparison. Incorporating ZeroSNR drags the mean value of StereoDiff's disparity maps closer to the GT, resulting in improved performance (Tab. 4).

even without per-scene optimization. In this work, we adopt MonST3R [91] as the stereo matching pipeline, which fine-tunes DUSt3R [72] with extensive dynamic video data. Compared to DUSt3R, MonST3R more accurately assigns zero confidence to potential low-quality correspondences (*e.g.*, dynamic, blurry) and applies SfM only to static, clear correspondences, significantly enhancing the performance and robustness in dynamic scenes. Typically, an optimization-based post-processing step is applied for improved global alignment after obtaining stereo matching results. However, we exclude this step for three reasons: 1) video depth estimation is a perception task, which is better to be inference-based; 2) the optimization step is both resource-intensive[2] and time-consuming[3]; and 3) Similar to DUSt3R [72] and MASt3R [36], MonST3R [91] inherently maintains global consistency through its closed-form global point cloud initialization, which uses a Minimum Spanning Tree (MST) to find the optimal path in the pairwise stereo matching graph with maximum confidence, followed by rigid point cloud registration [6, 43] to construct the final coarse 3D point clouds. As a result, StereoDiff is not only training-free but also fully inference-based[4].

We denote the depth maps estimated only based on stereo matching as $\mathcal{D}_\text{s} = \{D_t^\text{s}\}_{t=0}^{T-1} = \Theta_\text{s}(\mathcal{I})$ and those only generated by video depth diffusion as $\mathcal{D}_\text{d} = \{D_t^\text{d}\}_{t=0}^{T-1} = \Theta_\text{d}(x \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathcal{I})$. As illustrated in Fig. 3, the magnitude spectrum two error sequences measured using AbsRel and $(1-\delta 1)$ (please see Sec. 4.1.3 for specific definitions) are visualized. It is evident that $\mathcal{D}_\text{s}$ exhibits significantly lower low-frequency errors compared to $\mathcal{D}_\text{d}$, indicating strong global consistency. Conversely, $\mathcal{D}_\text{d}$ performs much better in high-frequency domains, which primarily represent the local consistency. These findings demonstrate the promising potential of leveraging the priors from video depth diffusion models to greatly enhance the local consistency of $\mathcal{D}_\text{s}$ while maintaining its original high-quality global consistency.

3.3. Video Depth Diffusion for Local Consistency

Formally, taking $\mathcal{D}_\text{s}$ as input, the video depth diffusion model produces the final video depth prediction, expressed as: $\mathcal{D}_\text{sd} = \{D_t^\text{sd}\}_{t=0}^{T-1} = \Theta_\text{d}(\mathcal{D}_\text{s}, \mathcal{I})$. In this paper, we adopt DepthCrafter [32], a fine-tuned SVD model using $\sim$20K video sequences, to perform a one-step denoising of $\mathcal{D}_\text{s}$. Note that only the pre-trained weight is adopted. Unlike SfM-based video depth estimation, which adheres to the "first principle", video depth diffusion models take a purely "data-driven" approach. These models are fine-tuned from

[2] It requires > 80GB of graphics memory for videos with $\geqslant 300$ frames at a resolution of 512×384, making it impractical for long videos.

[3] Processing a 200-frame video at 512×384 resolution with a 300-iteration optimization takes over 15 minutes on an NVIDIA A800 GPU.

[4] We omit the Weiszfeld algorithm [49] for focal length estimation, as it requires only 10 iterations and back-propagates gradients into a minimal $T \times 1$ matrix, where T is the number of frames.

[5] Comparisons are conducted in disparity space rather than true-depth space, because both DepthCrafter and StereoDiff represent their video depth estimations using disparity maps.

ICCV LongVid Foundations Workshop #5

pre-trained video generative models on large-scale video depth data, mapping the RGB video directly to video depth.

As shown in Fig. 3, the depth maps produced by video depth diffusion models $\mathcal{D}_\mathrm{d}$ significantly outperform those based on stereo matching $\mathcal{D}_\mathrm{s}$ in high-frequency domains. Particularly, Fig. 3c depicts the amplitude ratio of the error sequences calculated on $\mathcal{D}_\mathrm{s}$ and $\mathcal{D}_\mathrm{d}$ for clearer demonstration. This suggests that the components in higher frequency domains of $\mathcal{D}_\mathrm{s}$, which much more significantly differ from the GT distribution learned by the video depth diffusion models, are more likely treated as noise and effectively denoised. Conversely, the low-frequency characteristics of $\mathcal{D}_\mathrm{s}$ align much more closely with the GT video depth distribution, drawing less attention during denoising and thus being better preserved. This results in strong retention of low-frequency features and targeted denoising of high-frequency components, significantly reducing the high-frequency errors in $\mathcal{D}_\mathrm{s}$.

Mathematically, substituting $\mathcal{D}_\mathrm{s}$ into Eq. 1 yields the corresponding error sequence $\mathcal{E}_\mathrm{s} = \{\epsilon_t^\mathrm{s}\}_{t=0}^{T-1}$. This temporal signal can then be transformed into the frequency domain $\mathcal{F}(\epsilon_k^\mathrm{s}), k \in [0, T-1]$ using FFT (Eq. 2). Similarly, we denote the error sequence of $\mathcal{D}_\mathrm{sd}$ as $\mathcal{E}_\mathrm{sd} = \{\epsilon_t^\mathrm{sd}\}_{t=0}^{T-1}$. $\forall t \in [0, T-1]$, $\epsilon_t^\mathrm{s} \geqslant 0$ and $\epsilon_t^\mathrm{sd} \geqslant 0$. The average of error sequence yields the final metric: $(1/T)\sum_{t=0}^{T-1}\epsilon_t$. As discussed above and demonstrated in Fig. 3, during the second stage of StereoDiff, the video depth diffusion model acts as a "low-pass filter" on $\mathcal{F}(\epsilon_k^\mathrm{s})$. Assuming a threshold K_thr, for simplicity, we approximate that after the video depth diffusion process, the magnitudes of all frequency components $> K_\mathrm{thr}$ are re-scaled by a factor $\alpha \in (0, 1)$:

$$\mathcal{F}(\epsilon_k^\mathrm{sd}) \approx \begin{cases} \mathcal{F}(\epsilon_k^\mathrm{s}), & k \leqslant K_\mathrm{thr} \\ \alpha \cdot \mathcal{F}(\epsilon_k^\mathrm{s}), & k > K_\mathrm{thr} \end{cases} \tag{4}$$

Following Parseval's energy theorem, which states that the total energy of the signal in the time domain and frequency domain are equal, we can derive:

$$\begin{aligned} & \sum_{k=0}^{T-1} \left|\mathcal{F}(\epsilon_k^\mathrm{sd})\right|^2 \leqslant \sum_{k=0}^{T-1} \left|\mathcal{F}(\epsilon_k^\mathrm{s})\right|^2 \\ \Rightarrow & \sum_{t=0}^{T-1} \left|\epsilon_t^\mathrm{sd}\right|^2 \leqslant \sum_{t=0}^{T-1} \left|\epsilon_t^\mathrm{s}\right|^2 \Rightarrow \frac{1}{T}\sum_{t=0}^{T-1}\epsilon_t^\mathrm{sd} \leqslant \frac{1}{T}\sum_{t=0}^{T-1}\epsilon_t^\mathrm{s} \end{aligned} \tag{5}$$

This derivation shows that maintaining the low-frequency characteristics of $\mathcal{D}_\mathrm{s}$, while reducing the high-frequency components of its error sequence $\mathcal{E}_\mathrm{s}$, leads to improved performance. In practice, as illustrated in Fig. 3, StereoDiff's low-frequency error magnitudes $\mathcal{D}_\mathrm{sd}$ largely inherit those of $\mathcal{D}_\mathrm{s}$, while high-frequency components are significantly reduced by leveraging the video depth diffusion, leading to improved performance (Tab. 1 and 2) and greatly smoothed prediction (Fig. 1), aligning well with our analysis.

ZeroSNR. In diffusion models, the forward process progressively adds Gaussian noise to clean samples according to a pre-defined variance schedule, *i.e.*, $\beta_1, \cdots, \beta_T$:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right) \tag{6}$$

Let $\alpha_t = 1 - \beta_t$ and $\overline{a}_t = \prod_{s=1}^{t}\alpha_s$, x_t can be sampled as:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\overline{a}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}\right) \tag{7}$$

Equivalently:

$$x_t = \sqrt{\overline{a}_t}x_0 + \sqrt{1-\overline{a}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{8}$$

The SNR is defined as: $\mathrm{SNR}(t) = \overline{a}_t/(1-\overline{a}_t)$. Specifically, in DepthCrafter [32] and the standard SVD [4] scheduler[6], the variance sequence is $\beta_0 = 0.00085$ and $\beta_T = 0.012$ with linear scaling, we derive: $x_T \approx 0.0016x_0 + 0.9992\epsilon$. This indicates the input, *i.e.*, x_T, always contains a small amount of signal during training. The leaked signal contains the lowest frequency information, *e.g.*, the mean value. The model learns to denoise with this signal. However, during inference, pure Gaussian noise is used, prompting the model to generate outputs with medium value [40, 44].

As illustrated in Fig. 4, DepthCrafter's video disparity maps have a mean value closer to 0.5 compared to other methods. Although StereoDiff achieves relatively accurate mean disparity values without ZeroSNR due to its first stage (stereo matching), incorporating ZeroSNR further aligns the mean value of StereoDiff's disparity maps closer to GT, resulting in improved performance (Tab. 4).

4. Experiments

4.1. Experimental Settings

4.1.1. Implementation Details

In the first stage, we set $n = 2$ for forming image pairs, symmetrizing them before feeding them into the stereo matching pipeline. The Weiszfeld algorithm [49] is adopted for camera intrinsics, and Procrustes alignment [43] is used for solving camera poses. The maximum resolution is limited to 512. In the second stage, following [32], we set the window size to 110 frames with a 25-frame overlap. The ZeroSNR trick is implemented by setting the `trailing` [40, 44] mode for the timestep spacing in schedulers. Depth maps obtained from the first stage $\mathcal{D}_\mathrm{s}$ are resized to the original frame size using `nearest` interpolation before the one-step denoising process, which is performed from denoising timestep $t = 2$ to $t = 1$ with a total number of denoising timesteps $T = 4$.

[6] https://huggingface.co/docs/diffusers/en/api/schedulers/euler

Method	Bonn [47]			KITTI [24]			ScanNetV2 [13]			Sintel [8]			Average
	AbsRel ↓	RMSE ↓	$\delta 1$ ↑	AbsRel ↓	RMSE ↓	$\delta 1$ ↑	AbsRel ↓	RMSE ↓	$\delta 1$ ↑	AbsRel ↓	RMSE ↓	$\delta 1$ ↑	Rank ↓
DepthAnything V2 [81]	0.1250	1.7765	0.8297	0.1758	4.2583	0.6872	0.1445	0.2926	0.7808	0.3983	6.5771	0.5666	6.9
DepthAnything [80]	0.1112	1.5191	0.8860	0.1755	4.3756	0.6875	0.1409	0.2500	0.7978	0.3342	5.5025	0.5833	5.4
DUSt3R [72]	0.1757	2.3618	0.7798	0.3343	7.0966	0.5065	**0.0544**	**0.1184**	0.9782	1.9245	9.8570	0.3964	8.9
MASt3R [36]	0.1748	2.2829	0.7698	0.2250	5.0800	0.6460	0.0957	0.2251	0.9319	0.6130	4.7154	0.5063	7.9
MonST3R [91]	0.0818	1.2412	0.9542	0.1661	4.1881	0.7387	0.0907	0.1631	0.9162	0.5291	4.2812	0.5053	4.1
MonST3R$_{\textbf{OPT}}$ [91]	–	–	–	0.1635	**4.0935**	0.7496	–	–	–	0.5118	**4.2606**	0.5263	3.9
ChronoDepth [61]	0.1248	1.6918	0.8501	0.1749	4.4265	0.7288	0.1955	0.3198	0.6766	0.5421	4.3168	0.5286	7.2
DepthCrafter [32]	0.1104	1.6817	0.8955	0.1617	5.3883	0.7695	0.1879	0.4003	0.6650	**0.2861**	6.1423	**0.6972**	5.7
DepthAnyVideo [17]	0.0942	1.4982	0.9308	0.1487	5.3931	**0.8002**	0.1834	0.4202	0.6771	0.3363	5.5432	0.6378	5.3
StereoDiff$_{\text{DUSt3R}}$	0.1521	2.1402	0.7981	0.2600	6.7388	0.5661	0.0573	0.1393	**0.9789**	1.6521	7.8762	0.3848	8.2
StereoDiff$_{\text{MASt3R}}$	0.1491	2.0866	0.8126	0.1958	5.4359	0.6769	0.0989	0.2600	0.9358	0.4800	7.3534	0.5242	7.7
StereoDiff (Ours)	**0.0799**	**1.2257**	**0.9549**	**0.1469**	4.4183	0.7764	0.0944	0.1985	0.9060	0.3275	5.2812	0.5782	**2.9**

Table 1. **Quantitative comparison of StereoDiff with SoTA methods on zero-shot video depth benchmarks.** The five sections from top to bottom represent: image depth estimators, stereo matching-based estimators, video depth diffusion models, StereoDiff using other stereo matching methods, and StereoDiff. To make sure a comprehensive evaluation, we used four datasets: Bonn [47], KITTI [24], ScanNetV2 [13], and Sintel [8]. We report the mean metrics of StereoDiff across 10 independent runs. MonST3R$_{\textbf{OPT}}$ (OPT: with optimization) can not be evaluated on long video depth benchmarks (*i.e.*, Bonn and ScanNetV2) due to computational constraints, please see footnote 2 and 3 for more details. Best results are **bolded** and the second best are underlined.

4.1.2. Evaluation Datasets.

We validate StereoDiff on four *zero-shot* video depth benchmarks: Bonn [47], KITTI [24], ScanNetV2 [13], and Sintel [8]. 6 dynamic indoor videos from Bonn (with 332 $\sim$ 580 frames each) are used. 12 dynamic outdoor videos from KITTI's validation set (with 17 $\sim$ 251 frames each) are used. 23 dynamic synthetic scenes (each contains 20$\sim$50 frames) of Sintel are used. Randomly selected 4 scenes (`scene0078_00`, `scene0192_01`, `scene0348_00`, `scene-0556_01`) in ScanNetV2 are used. The frame resolution of Bonn, KITTI, ScanNetV2, Sintel are resized to 640×480, 1216×352, 1024×448, 1280×960, respectively.

4.1.3. Evaluation Metrics.

Following the affine-invariant evaluation protocols from [17, 19, 27, 32, 33, 44, 61, 91], we firstly align the estimated video depth maps with GT using least-squares fitting, and resize all estimations to match the original size of input video in `nearest` mode. Note that during the least-squares fitting, all frames in a video depth sequence share *identical* scaling and shifting factors, same as DepthCrafter [32] and MonST3R [91]. *Temporal inconsistencies* will lead to worse metrics, *e.g.*, testing MonST3R on Bonn with *per-frame* scale and shift yields an AbsRel of 0.0341, much better than the reported 0.0818. Specifically, given GT $\mathcal{D}^* = \{D_t^*\}_{t=0}^{T-1}$ and fitted predictions $\hat{\mathcal{D}} = \{\hat{D}_t\}_{t=0}^{T-1}$, we report two error metrics: 1) absolute mean relative error (AbsRel) and 2) root-mean-square deviation (RMSE), *i.e.*:

$$
\begin{aligned}
\text{AbsRel}(\mathcal{D}^*, \hat{\mathcal{D}}) &= \frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{1}{N}\sum_{j=0}^{N-1}\frac{|D_{tj}^* - \hat{D}_{tj}|}{\hat{D}_{tj}}\right] \\
\text{RMSE}(\mathcal{D}^*, \hat{\mathcal{D}}) &= \frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{1}{N}\sqrt{\sum_{j=0}^{N-1}(D_{tj}^* - \hat{D}_{tj})^2}\right]
\end{aligned} \tag{9}
$$

where $N = H \times W$, indicating the total number of pixels. We also report one accuracy metric: $\delta 1$, denoting the proportion of pixels satisfying $\text{Max}(D_{tj}^*/\hat{D}_{tj}, \hat{D}_{tj}/D_{tj}^*) < 1.25$.

4.2. Quantitative Comparisons

As shown in Tab. 1, StereoDiff achieves the *best* comprehensive results across four zero-shot video depth benchmarks. Furthermore, the results of frequency domain analysis (Tab. 2) demonstrate that StereoDiff effectively maintains the strong low-frequency global consistency achieved via stereo matching, while significantly enhancing the high-frequency local consistency. This enhancement greatly reduces local jitters and flickering across neighboring frames particularly in dynamic areas (Fig. 1), as high-frequency characteristics of $\mathcal{D}_s$ differ much more significantly from the GT distribution learned by the video depth diffusion models, and are more likely treated as noise and effectively denoised. Additionally, Tab. 3 clearly shows that StereoDiff outperforms MonST3R mainly in high-frequency dynamic regions and outperforms DepthCrafter mainly in low-frequency static regions. These results align well with our analysis in Sec. 3.2 and 3.3.

Metrics	Method	Low Freq. $\mathcal{F}_0$	$\mathcal{F}_1$	$\mathcal{F}_2$	$\mathcal{F}_3$	$\mathcal{F}_4$	$\mathcal{F}_5$	$\mathcal{F}_6$	$\mathcal{F}_7$	$\mathcal{F}_8$	$\mathcal{F}_9$	High Freq. $\mathcal{F}_{10}$
AbsRel↓	DepthCrafter	0.1104	0.0152	0.0215	0.0238	0.0286	0.0206	**0.0112**	**0.0062**	**0.0023**	**0.0012**	**0.0009**
	MonST3R	0.0822	**0.0130**	0.0149	0.0142	**0.0149**	**0.0142**	0.0144	0.0116	0.0077	0.0062	0.0067
	StereoDiff (Ours)	**0.0806**	0.0159	**0.0128**	**0.0132**	0.0157	0.0143	0.0135	0.0098	0.0067	0.0043	0.0032
RMSE↓	DepthCrafter	1.6823	0.1783	0.3221	0.2269	0.3125	0.2567	**0.1448**	**0.0884**	**0.0355**	**0.0191**	**0.0144**
	MonST3R	1.2427	**0.0949**	0.1075	**0.1633**	**0.1503**	0.1579	0.1604	0.1356	0.0848	0.0678	0.0726
	StereoDiff (Ours)	**1.2294**	0.1349	**0.1065**	0.1657	0.1659	**0.1565**	0.1469	0.1187	0.0786	0.0517	0.0421
$(1-\delta 1)$↓	DepthCrafter	0.1046	0.0380	0.0655	0.0696	0.0835	0.0619	**0.0331**	**0.0198**	**0.0100**	**0.0046**	**0.0027**
	MonST3R	0.0481	**0.0207**	0.0247	**0.0313**	**0.0408**	**0.0411**	0.0335	0.0258	0.0180	0.0134	0.0150
	StereoDiff (Ours)	**0.0478**	0.0246	**0.0241**	0.0325	0.0428	0.0442	0.0371	0.0261	0.0173	0.0101	0.0069

Table 2. **Quantitative comparisons on Bonn of MonST3R, DepthCrafter, and StereoDiff on different frequency domains.** We use DFT and Inverse DFT to disentangle the components of the metric sequences in various frequency domains. For simplicity, the entire frequency range is divided exponentially into 11 discrete groups: $\mathcal{F}_0, \cdots, \mathcal{F}_{10}$, representing low to high frequencies. We report the results on three well-recognized metrics, AbsRel ↓, RMSE ↓, and $(1-\delta 1)$ ↓. Please see the results on KITTI in the Supplementary Materials.

Region	AbsRel ↓	RMSE ↓	$\delta 1$ ↑	$\delta 2$ ↑
Dynamic	**-0.0069**	**-0.0844**	**+0.0140**	**+0.0023**
Overall	-0.0020	-0.0150	+0.0013	-0.0042
Static	+0.0009	0	-0.0004	-0.0049

(a) Performance improvement of StereoDiff over MonST3R. For example, AbsRel $=$ AbsRel$_{\text{StereoDiff}}$ $-$ AbsRel$_{\text{MonST3R}}$.

Region	AbsRel ↓	RMSE ↓	$\delta 1$ ↑	$\delta 2$ ↑
Dynamic	-0.0178	-0.2575	+0.0413	**-0.0055**
Overall	-0.0306	-0.4555	+0.0600	-0.0071
Static	**-0.0335**	**-0.4990**	**+0.0641**	-0.0069

(b) Performance improvement of StereoDiff over DepthCrafter.

Table 3. **Quantitative comparisons on dynamic and static regions of Bonn among MonST3R, DepthCrafter and StereoDiff.** We use FlowSAM [78] for masking moving areas. Please see the results on KITTI in the Supplementary Materials.

Method	AbsRel↓	RMSE↓	$\delta 1$↑	$\delta 2$↑
Naive Solution *w/* Latent Sharing *w/o* ZeroSNR	0.1245 ±0.0002	1.7807 ±0.0016	0.8503 ±0.0018	0.9719 ±0.0006
w/o Latent Sharing *w/o* ZeroSNR	0.0809 ±0.0003	1.2383 ±0.0039	0.9544 ±0.0006	0.9867 ±0.0003
StereoDiff (Ours) *w/o* Latent Sharing *w/* ZeroSNR	**0.0799** ±0.0001	**1.2257** ±0.0028	**0.9549** ±0.0006	**0.9870** ±0.0004

Table 4. **Ablation studies.** Removing latent sharing strategy and adding the ZeroSNR trick both yield effective performance gains. Here we report the results on Bonn dataset.

5. Ablation Study

As discussed in Sec. 2.2, for video diffusion-based video depth estimators, input videos are typically divided into windows and processed sequentially. In DepthCrafter, this is performed by dividing the video into overlapped windows and sharing the latents of overlapped frames. While this strategy improves continuity, it can still fall short in maintaining consistency between windows, especially on static backgrounds (Fig. 1). As illustrated in Tab. 4, the removal of latent sharing strategy leads to significant performance gains. This is primarily because: 1) the strict spatial correspondence between the diffusion's latent space and the RGB space, making latent sharing ineffective for scenes with moving cameras or objects, which may lead to harmful feature distortions, especially as the timestep $t \to 0$; and 2) in DepthCrafter's original multi-step denoising process, the latent is progressively refined from Gaussian noise, where sharing latents across overlapping frames can not only aids consistency at early timesteps ($t \to T$) but also allows the distortions of latent feature to be gradually refined as $t \to 0$. Additionally, incorporating ZeroSNR aligns the mean value of StereoDiff's disparity maps more closely with the GT (Fig. 4), further enhancing the performance.

6. Conclusion

Conclusion. In this paper, we emphasize the need for distinct strategies to achieve consistent video depth estimation across static and dynamic regions. Motivated by these insights, we introduce StereoDiff, a novel two-stage video depth estimator that combines stereo matching for strong global consistency provided by the global 3D constraints, and video depth diffusion for significantly enhanced local consistency. Experimental results on two well-acknowledged video depth benchmarks (Tab. 1), including the frequency domain analysis (Fig. 3, Tab. 2), demonstrate StereoDiff's effectiveness in synergizing the strengths of both, achieving SoTA performance in dynamic, zero-shot, real-world video depth estimation.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot trans-

fer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3

[2] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[3] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Tat-Jin Chin, Chunhua Shen, and Ian Reid. Auto-rectify network for unsupervised indoor depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 3

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 3, 6

[5] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 3

[6] Romain Brégier. Deep regression on manifolds: a 3D rotation case study. In *2021 International Conference on 3D Vision (3DV)*, 2021. 5

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 3, 2024. 3

[8] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012. 2, 7

[9] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8001–8008, 2019. 3

[10] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 3

[11] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 3

[12] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7063–7072, 2019. 3

[13] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 7

[14] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 3

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 3

[16] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 3

[17] H. Yang etc. Depth any video with scalable synthetic data. *arXiv*, 2024. 2, 3, 7

[18] X. Luo etc. Consistent video depth estimation. *SIGGRAPH*, 2020. 3

[19] X. Zhang etc. Betterdepth: Plug-and-play diffusion refiner for zero-shot monocular depth estimation. *NeurIPS*, 2024. 2, 3, 7

[20] Z. Zhang etc. Consistent depth of moving objects in video. *SIGGRAPH*, 2021. 3

[21] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 3

[22] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2025. 2, 3

[23] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 1272–1279, 2013. 2, 3, 4

[24] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 2, 7

[25] Songen Gu, Wei Yin, Bu Jin, Xiaoyang Guo, Junming Wang, Haodong Li, Qian Zhang, and Xiaoxiao Long. Dome: Taming diffusion model into high-fidelity controllable occupancy world model. *arXiv preprint arXiv:2410.10429*, 2024. 3

[26] Jing He, Haodong Li, Yongzhe Hu, Guibao Shen, Yingjie Cai, Weichao Qiu, and Ying-Cong Chen. Disenvisioner: Disentangled and enriched visual prompt for customized image generation. *arXiv preprint arXiv:2410.02067*, 2024. 3

[27] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Liu, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. *arXiv preprint arXiv:2409.18124*, 2024. 2, 3, 7

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[29] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[30] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3

[31] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. 3

[32] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 1, 2, 3, 5, 6, 7

[33] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 2, 3, 7

[34] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1621, 2021. 3

[35] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 3

[36] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 2, 3, 4, 5, 7

[37] Xiao-Lei Li, Haodong Li, Hao-Xiang Chen, Tai-Jiang Mu, and Shi-Min Hu. Discene: Object decoupling and interaction modeling for complex scene generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[38] 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.

[39] Yixun Liang*, Xin Yang*, Jiantao Lin, Haodong Li, Xiaogang Xu, and Ying-Cong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6517–6526, 2024. 3

[40] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5404–5411, 2024. 6

[41] Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G Narasimhan, and Jan Kautz. Neural rgb (r) d sensing: Depth and uncertainty from a video camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10986–10995, 2019. 3

[42] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, 39(4):71–1, 2020. 3

[43] Priyanka Mandikal, KL Navaneet, Mayank Agarwal, and R Venkatesh Babu. 3d-lmnet: Latent embedding matching for accurate and diverse 3d point cloud reconstruction from a single image. *arXiv preprint arXiv:1807.07796*, 2018. 5, 6

[44] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. *arXiv preprint arXiv:2409.11355*, 2024. 2, 3, 6, 7

[45] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[46] Kemal E Ozden, Konrad Schindler, and Luc Van Gool. Multibody structure-from-motion in practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1134–1141, 2010. 2, 3, 4

[47] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7855–7862. IEEE, 2019. 2, 5, 7

[48] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 3

[49] Frank Plastria. The weiszfeld algorithm: proof, amendments, and extensions. *Foundations of location analysis*, pages 357–389, 2011. 5, 6

[50] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3

[52] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 3

[53] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695, 2022. 2, 3

[55] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[56] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 3

[57] Mohamed Sayed, Filippo Aleotti, Jamie Watson, Zawar Qureshi, Guillermo Garcia-Hernando, Gabriel Brostow, Sara Vicente, and Michael Firman. Doubletake: Geometry guided depth estimation. *arXiv preprint arXiv:2406.18387*, 2024. 3

[58] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 2, 3, 4

[59] 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*, 2016. 2, 3, 4

[60] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[61] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. *arXiv preprint arXiv:2406.01493*, 2024. 2, 3, 7

[62] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 3, 4

[63] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

[64] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[65] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15598–15607, 2021. 3

[66] Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 3

[67] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. *arXiv preprint arXiv:1812.04605*, 2018. 3

[68] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 3

[69] George Vogiatzis and Carlos Hernández. Video-based, real-time multi-view stereo. *Image and vision computing*, 29(7): 434–441, 2011. 3

[70] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *2019 International Conference on 3D Vision (3DV)*, pages 348–357. IEEE, 2019. 3

[71] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 2, 4

[72] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 4, 5, 7

[73] Yiran Wang, Zhiyu Pan, Xingyi Li, Zhiguo Cao, Ke Xian, and Jianming Zhang. Less is more: Consistent video depth estimation with masked frames modeling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6347–6358, 2022. 3

[74] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 3

[75] Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9466–9476, 2023. 2, 3

[76] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 2, 4

[77] Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, pages 127–134. IEEE, 2013. 3, 4

[78] Junyu Xie, Charig Yang, Weidi Xie, and Andrew Zisserman. Moving object segmentation: All you need is sam (and flow). *arXiv preprint arXiv:2404.12389*, 2024. 8

[79] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 3

[80] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 3, 7

[81] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 3, 7

[82] Rajeev Yasarla, Hong Cai, Jisoo Jeong, Yunxiao Shi, Risheek Garrepalli, and Fatih Porikli. Mamo: Leveraging memory and attention for monocular video depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8754–8764, 2023. 2, 3

[83] Rajeev Yasarla, Manish Kumar Singh, Hong Cai, Yunxiao Shi, Jisoo Jeong, Yinhao Zhu, Shizhong Han, Risheek Garrepalli, and Fatih Porikli. Futuredepth: Learning to predict the future improves video depth estimation. In *European Conference on Computer Vision*, pages 440–458. Springer, 2025. 2, 3

[84] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020. 3

[85] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021. 3

[86] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 3

[87] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 3

[88] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3916–3925, 2022. 3

[89] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. *Advances in Neural Information Processing Systems*, 35:14128–14139, 2022. 3

[90] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1725–1734, 2019. 3

[91] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 1, 2, 3, 4, 5, 7

[92] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[93] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022. 3

[94] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 3

[95] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 3