StableDepth: Scene-Consistent and Scale-Invariant Monocular Depth

Zheng Zhang[1,2,⋆,*] Lihe Yang[1,⋆] Tianyu Yang[2,4,†] Chaohui Yu[2,4]
Xiaoyang Guo[3] Yixing Lao[1] Hengshuang Zhao[1,†]
[1]The University of Hong Kong [2]DAMO Academy, Alibaba Group
[3]The Chinese University of Hong Kong [4]Hupan Lab
https://stabledepth.github.io

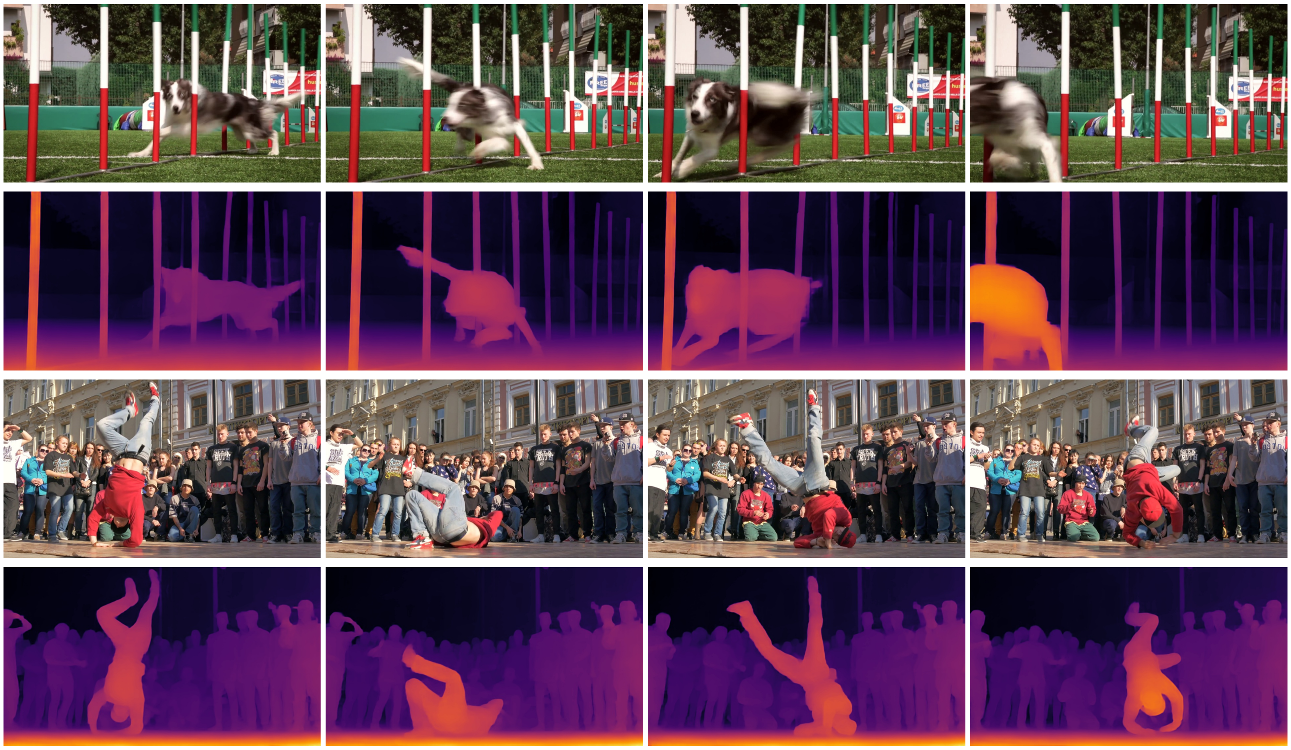

Figure 1. StableDepth achieves efficient online monocular depth estimation that produces scene-consistent and scale-invariant predictions frame by frame. As shown in the above challenging scenarios, StableDepth generates stable depth maps with precise geometric details and effectively handles dynamic motions, without the need of accessing future frames.

Abstract

Recent advances in monocular depth estimation significantly improve robustness and accuracy. However, relative depth models exhibit flickering and 3D inconsistency in video data, limiting 3D reconstruction applications. We introduce StableDepth, a scene-consistent and scale-invariant depth estimation method achieving scene-level 3D consistency. Our dual-decoder architecture learns from large-scale unlabeled video data, enhancing generalization and reducing flickering. Unlike previous methods requiring full video sequences, StableDepth enables online inference at 13× faster speed, achieving significant improvements across benchmarks with comparable temporal consistency to video diffusion-based estimators.

1. Introduction

Monocular Depth Estimation (MDE) [12, 45] is pivotal in bridging the gap between 2D and 3D data representation, serving critical roles in a wide range of downstream applica-

⋆Joint first authors. *Work done when interning at DAMO Academy.
† Corresponding author.

tions, such as autonomous driving [51], robotics [54], medical imaging [32], and virtual reality [13]. Advancements in foundational models [35, 43] have greatly enhanced the accuracy and robustness of MDE [25, 56]. Despite these advances, the intrinsic ambiguity of monocular cues still poses substantial challenges, particularly for downstream applications that require 3D-consistent data and maintaining stability for uncontrolled in-the-wild data.

MDE is primarily categorized into two main task settings: relative depth estimation and metric depth estimation. Relative depth estimation, as explored in studies [15, 18, 25, 39, 56], involves predicting depth values within a normalized depth range of [0, 1]. This scale- and shift-invariant regression target facilitates robust joint training across diverse data domains. However, the predicted depth values differ from the actual depth due to scale and shift discrepancies, leading to inconsistency across video frames. This inconsistency can complicate advanced downstream tasks like 3D reconstruction. In contrast, metric depth estimation [4, 22, 37, 59] aims to predict physical distances in metric scale. However, these approaches face significant challenges. Scaled versions of a scene's depth values might all represent plausible depth estimations of the actual depth. Such ambiguity complicates the learning process of depth regression, hindering both generalization capabilities and precise 3D structures. Severe flickering and 3D inconsistency are also observed given video inputs. This challenge prompts us to reconsider *what is an effective representation of MDE foundation models*.

The limitations of existing approaches are summarized in Tab. 1, which compares different depth estimation paradigms. While relative depth models [25, 39, 56] offer good generalization across datasets but suffer from per-frame inconsistency, metric depth models [4, 59] provide real-world measurements but struggle with generalization across diverse scenes. Scale-invariant models [20] improve upon relative models by eliminating shift ambiguity while maintaining scale invariance, but they still lack consistency across frames in the same scene.

To address these challenges, we propose that the ideal task setting should be scene-consistent and scale-invariant (SCSI) depth estimation, as shown in Tab. 1. Our approach strikes a balance between per-frame scale-invariant methods that lack consistency and purely metric approaches that struggle with generalization. Specifically, our model primarily generates scale-invariant depth maps, but in the same scene or video, predictions across multiple frames demonstrate 3D consistency. Additionally, the scale factors for multi-frame data remain consistent compared with the actual depth values. In this work, the development of our model is primarily guided by the following objectives: 1) ensuring high-quality, *scene-consistent, and scale-invariant* depth estimations to achieve robust video consistency; 2) enhancing strong generalization capabilities and precise depth details.

Depth Type	Scale	Shift	General	Recons	Consist
Relative	Per-frame	Per-frame	✓	✗	✗
Scale-invariant	Per-frame	None	✓	✓	✗
Metric	None	None	✗	✓	✗
SCSI (Ours)	Per-scene	None	✓	✓	✓

Table 1. Comparison of different depth types. Our proposed SCSI depth combines the strengths of relative and metric depth while overcoming their weaknesses. **General**: generalization capability. **Recons**: reconstruction quality. **Consist**: scene consistency.

Our framework introduces three key innovations to achieve scene-consistent depth estimation. First, we leverage temporal priors from pre-trained video diffusion models to generate pseudo-labels with inherent 3D consistency, replacing traditional unstable photometric constraints [3, 23, 47, 48]. Second, we design a dual-branch architecture where one branch processes labeled data for metric accuracy, while the other aligns unlabeled video frames to pseudo-labels through scene-level normalization, decoupling conflicting objectives. Third, our semi-supervised paradigm jointly optimizes metric supervision and temporal alignment, enabling the model to generalize across domains while eliminating flickering artifacts. This unified approach supports real-time inference without requiring future frames, overcoming the limitations of both single-frame and video-based methods.

We evaluate our model on multiple datasets under zero-shot settings, aligning each scene using only a single scale (SCSI). As demonstrated in Figure 1, our method robustly handles challenging scenarios including dynamic motions and complex textures while preserving geometric fidelity frame-by-frame. Compared to DepthCrafter [23], we achieve 13× faster inference speed with online processing capability, improving prediction accuracy on Sintel [5], Bonn [36], ScanNet v2 [9], and KITTI [17] by 13.2%, 86.8%, 39.3%, and 8.2%, respectively. Our contributions are summarized as follows:

- We introduce StableDepth, a novel monocular depth estimation framework that achieves scene-consistent and scale-invariant (SCSI) depth estimation, bridging the gap between relative and metric depth prediction while maintaining temporal consistency in video sequences.
- We propose a video diffusion baking strategy that leverages pre-trained video diffusion models for generating high-quality pseudo-labels, enabling effective semi-supervised learning from unlabeled video data while preserving temporal consistency.
- We develop a dual-decoder architecture that simultaneously handles supervised metric depth estimation and unsupervised depth alignment, effectively addressing the target misalignment issue while maintaining accuracy.

- We demonstrate state-of-the-art performance across multiple benchmarks, achieving up to 86.8% improvement in accuracy while providing 13× faster inference compared to previous methods. Our method enables efficient, online inference while maintaining comparable temporal stability to video diffusion models.

2. Related Work

Monocular depth estimation. MDE focuses on inferring depth from a single image. Early works [8, 31, 55] learn ordinal relationships from coarse annotations but cannot recover geometric structure. Recent methods [11, 39, 56, 57] leverage large-scale data to learn affine-invariant depth with improved generalization. However, unknown depth shifts cause geometric distortions, hindering 3D applications. To address this, recent works attempt zero-shot metric depth estimation [4, 19, 22, 37, 38, 59, 61] and explore self-supervised scale-consistent learning [50, 60], though maintaining consistent scale across scenes remains challenging.

Towards consistent video depth. Video depth estimation requires both temporal consistency and per-frame precision. Traditional approaches use post-processing optimization [27, 34] or temporal modeling [29, 49], but incur computational costs or fail in dynamic scenes. Other methods leverage camera poses [46, 53] or memory-based approaches [52, 58], but often sacrifice geometric fidelity for smoothness. Recent works like Video Depth Anything [7] and video-free approaches [26] show promise for temporal consistency in long sequences. Video Depth Anything freezes the Depth Anything V2 encoder and incorporates cross-frame attention in the DPT [40] decoder for temporal consistency, though these methods still produce depth with inherent shift ambiguity.

Diffusion priors for depth modeling. Video diffusion models [21, 24] demonstrate potential for consistent depth estimation. DepthCrafter [23] achieves temporal stability but generates only relative depths, while ChronoDepth [47] suffers from limited temporal contexts. Our proposed SCSI depth estimation leverages video diffusion baking and dual-decoder architecture for temporally consistent depth suitable for efficient online inference.

3. StableDepth

Our work leverages both labeled images and unlabeled video data to enhance SCSI depth predictions, achieving video consistency and enhanced generalization capabilities through the distillation of video diffusion priors. Formally, we denote the sets of labeled image-depth pairs and unlabeled videos as $\mathcal{D}^l = \{(x_i, d_i)\}_{i=1}^M$ and $\mathcal{D}^u = \{v_i\}_{i=1}^N$, respectively. Our objective is to utilize a high-quality labeled image dataset, Hypersim [42], to help the model grasp real-world scale, while employing unlabeled video data to guide the model's depth predictions toward achieving video consistency and improved generalization.

3.1. Mining Consistency from Unlabeled Videos

StableDepth is designed to learn from both labeled images and unlabeled videos. For labeled image data, we predict the depth in metric space. The activation function of the network's last layer is sigmoid, after which its predictions are scaled to match the dataset values by multiplying with a predefined maximum depth. The SiLogLoss [12, 44, 57] is then used to minimize the error between the prediction and the ground truth:

$$\mathcal{L}_{\text{label}} = \sqrt{\frac{1}{HW}\sum_{i=1}^{HW}\left(\log\frac{d_i}{d_i^*}\right)^2 - \lambda\left(\frac{1}{HW}\sum_{i=1}^{HW}\log\frac{d_i}{d_i^*}\right)^2}, \tag{1}$$

where d_i^* and d_i are the prediction and ground truth, respectively, and λ is a hyper-parameter.

For unlabeled video data, we leverage video diffusion models, specifically DepthCrafter [23], to generate pseudo-labels due to their inherent advantages in maintaining temporal consistency. These models, pre-trained on large-scale video data with cross-frame attention mechanisms, can generate stable depth predictions across entire video segments while preserving a consistent global scale and shift relationship with metric depth. Unlike conventional image-based MDE approaches that process frames independently, video diffusion models account for temporal dependencies, making them particularly effective for generating reliable pseudo-labels that exhibit strong video consistency - a crucial property for our scene-consistent training objective. For a video, we need a method to normalize the depth predicted by our model to the same space as the pseudo-labeled depth. Unlike the normalization used in single-frame relative depth training, this approach focuses on ensuring that an entire video has a unique scale and shift, rather than each frame having an independent scale and shift. This method aims to achieve video consistency in the depth predictions within the same scene. To supervise our model, we first use the following formula to normalize the pseudo-labels,

$$\hat{D}(v_i) = \frac{D(v_i) - m(D(v_i))}{s(D(v_i))}, \tag{2}$$

where $D(v_i)$ is the pseudo-label for video v_i, and $m(D(v_i))$ and $s(D(v_i))$ are used to remove the shift and scale for this video, respectively:

$$\begin{aligned} m(D(v_i)) &= \text{median}(D(v_i)), \\ s(D(v_i)) &= \frac{1}{THW}\sum_{i=1}^{THW}\left|D(v_i) - m(D(v_i))\right|. \end{aligned} \tag{3}$$

For the frame-by-frame depth predictions of video v_i from our model, we concatenate them to obtain $P(v_i)$ with

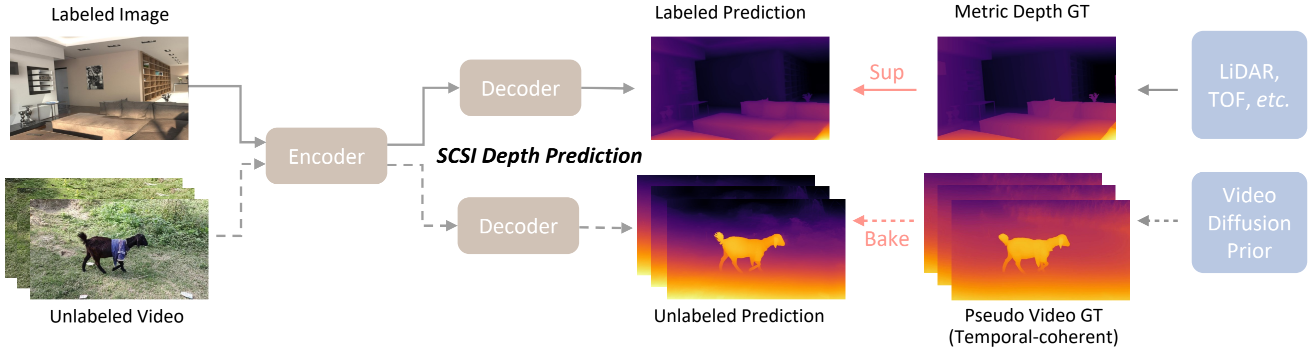

Figure 2. **Overview of StableDepth.** StableDepth employs a shared encoder with dual decoders to process both labeled images and unlabeled videos. The upper branch handles labeled images with direct metric depth supervision ("Sup"), while the lower branch processes unlabeled video sequences using temporal-coherent pseudo-labels generated by pre-trained video diffusion priors [23] ("Bake"). This architecture enables scene-consistent and scale-invariant (SCSI) depth prediction by combining the benefits of metric depth supervision with temporal consistency learning from video sequences. Solid arrows indicate the primary data flow, while dashed arrows represent the auxiliary training path for unlabeled data.

dimensions $[T, H, W]$. We then apply the same normalization process to the depth predictions $P(v_i)$ to obtain its normalized version $\hat{P}(v_i)$. For $\hat{P}(v_i)$ and $\hat{D}(v_i)$, we use the following alignment loss to minimize their difference,

$$\mathcal{L}_{\text{unlabel}} = \frac{1}{THW} \sum_{i=1}^{THW} \left| \hat{P}(v_i) - \hat{D}(v_i) \right|. \tag{4}$$

In one iteration, we simultaneously sample a certain proportion of labeled images and unlabeled videos for joint training, and use the following combined loss as the final loss for this iteration:

$$\mathcal{L} = \mathcal{L}_{\text{label}} + \gamma \mathcal{L}_{\text{unlabel}}, \tag{5}$$

where γ gradually increases following a cosine curve as the number of iterations progresses.

3.2. Avoiding Conflict in Depth Labels

In the field of video consistency for depth prediction, most research [23] focuses on improving video consistency in near-field scenes. Therefore, we use depth in the disparity space for video data training.

Following the previous network structure [2, 39, 56, 57], we use the ViT-L encoder [10], denoted as $\mathcal{E}$, paired with the decoder from DPT [40], denoted as $\mathcal{D}_1$. The decoder outputs depth in SCSI space with a range of (0, 1). A natural approach is to utilize this original decoder to predict the disparity-space depth of a video sequence, and then take the reciprocal of the predictions. Specifically, the depth prediction $P(v_i)$ is given by:

$$P(v_i) = \frac{1}{\mathcal{D}_1(\mathcal{E}(v_i))}. \tag{6}$$

Finally, the depth predictions are normalized and supervised in semi-supervised settings following Eq. 3 and 4.

Unfortunately, in our experiments, we found that although this modification improves the video consistency of depth predictions for videos, it reduces the accuracy of depth prediction by the model, resulting in lower metrics such as δ_1. We speculate that this is mainly due to the used pseudo-labeling model [23], which, despite having good video consistency, still lags behind Depth Anything V2 [57] in terms of depth prediction accuracy, particularly in the prediction of distant objects. To address this issue, we propose to use an auxiliary network to decouple the prediction and supervision of labeled images and unlabeled videos. To this end, we experimented with two approaches:

In the first approach, both labeled images and unlabeled videos share the encoder $\mathcal{E}$ and the decoder $\mathcal{D}_1$. A lightweight CNN network, denoted as $\mathcal{C}$, is used to directly process the prediction results of the original decoder $\mathcal{D}_1$, with the CNN network outputting depth in disparity space:

$$P(v_i) = \mathcal{C}(\mathcal{D}_1(\mathcal{E}(v_i))). \tag{7}$$

In the second approach, labeled images and unlabeled videos share the encoder $\mathcal{E}$. An additional DPT [40] decoder, denoted as $\mathcal{D}_2$, is used to process the features extracted by the encoder $\mathcal{E}$ for unlabeled videos, as illustrated by Fig. 2. $\mathcal{D}_2$ directly produces depth in disparity space, while $\mathcal{D}_1$ continues to output depth in SCSI space.

$$P(x_i) = \mathcal{D}_1(\mathcal{E}(x_i)), \quad P(v_i) = \mathcal{D}_2(\mathcal{E}(v_i)). \tag{8}$$

The final results indicate that sharing the encoder between labeled images and unlabeled videos, while using two independent DPT decoders, achieves the best accuracy and video consistency.

3.3. Scene-Consistent Depth Evaluation

Scale alignment. For zero-shot evaluation, we enforce *single-scale consistency per scene* through RANSAC-based alignment with zero shift constraint. Each scene's predictions are scaled by a unified factor determined through 1,000 sampling iterations, retaining inliers with $\max(\frac{d_{pred}}{d_{gt}}, \frac{d_{gt}}{d_{pred}}) < \tau$, where $\tau = 1.25$. The optimal scale is refined via least-squares optimization over all inliers, preserving geometric fidelity while eliminating cross-frame scale ambiguity. This *per-scene single-scale paradigm* ensures coherent depth relationships across dynamic sequences without per-frame adjustments.

Motion-aware temporal metric. We propose the *Motion-aware Temporal Difference (MTD)* to quantify depth consistency in dynamic regions:

$$\text{MTD} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{\sum_{x,y} |\Delta D_t| \cdot W_t}{H \times W}, \tag{9}$$

where $\Delta D_t = D_t - D_{t-1}$ and $W_t(x,y) = (1 + \|\mathbf{f}_t\|_2)^{-1}$ weights the depth variations by optical flow magnitude $\mathbf{f}_t$ (computed via Farneback [14] algorithm). Lower MTD indicates better temporal stability, penalizing flickering in high-motion areas.

3.4. Discussion

In this section, we address several important aspects of our approach to provide further clarity on methodology, design choices, and relationships to existing work.

Decouple accuracy and consistency via dual learning. *Our dual-decoder architecture decouples depth accuracy from temporal consistency learning*, enabling joint optimization from diverse data sources. The primary decoder specializes in per-pixel depth precision through supervised training on high-quality labeled images, while the auxiliary decoder learns scene-level coherence from unlabeled videos via pseudo-labels. This separation resolves the inherent conflict where video-based methods like DepthCrafter [23] sacrifice depth accuracy for temporal smoothness, and single-frame approaches fail to ensure cross-frame consistency. Crucially, our design preserves absolute depth scales critical for 3D reconstruction, unlike relative-depth representations in video diffusion models.

Beyond video diffusion priors. We repurpose video diffusion models as *sources for scene consistency* rather than as direct depth estimators. While models like DepthCrafter [23] excel at generating temporally consistent sequences, they produce relative depth representations with limited accuracy, particularly for distant objects, making them unsuitable for downstream tasks requiring precise geometry. Our approach extracts the temporal consistency knowledge from these models through scene-level normalization while maintaining superior depth accuracy. This strategy enables online inference 13× faster than video-based methods while achieving both better depth precision and comparable temporal stability.

4. Experiment

4.1. Implementation Details

Network parameters and training details. We use the network architecture from Depth Anything V2 (DAv2) [57], consisting of a pre-trained ViT-L [10] encoder for feature extraction and a DPT [40] decoder for depth regression. The ViT-L encoder is initialized with parameters from the DAv2 model, while the primary decoder is randomly initialized with Sigmoid activation for metric depth output. The auxiliary decoder, also initialized randomly, uses ReLU activation for disparity (inverse depth) prediction. Each batch consists of a 1:2 ratio of labeled images to unlabeled video frames, totaling 24 frames. The learning rate for the encoder is 2.5×10^{-6}, while the decoder's learning rate is 10 times larger. We use the AdamW optimizer [33] with a weight decay of 0.01 and employ a polynomial decay schedule for the learning rate. For data augmentation, we apply horizontal flipping to labeled images. In Eq. (5), the maximum value of γ is set to 1. In Eq. (1), λ is set to 0.5.

Training dataset and pseudo-labeling. To train our model, we utilize DepthCrafter [23], a model finetuned on SVD [3], to generate video-consistent pseudo-labels for unlabeled video data. For metric depth supervision, we use Hypersim [42], a high-quality synthetic dataset with ∼60K frames, as training data for decoder $\mathcal{D}_1$. We select 6,000 videos from SA-V [41], which features dynamic foreground objects, yielding approximately ∼200K frames for pseudo-labeling training. To enhance the model's performance on distant scenes where DepthCrafter shows limitations, we incorporate additional datasets with accurate depth labels including VKITTI2 [6, 16], LightwheelOcc [30], and MatrixCity [28], treating frames from the same scene as video sequences during training.

4.2. Evaluation

Evaluation datasets. We evaluate on four datasets spanning diverse scenarios: Sintel [5] (23 synthetic videos, 50 frames each with precise depth labels), Bonn [36] (26 dynamic videos using frames 30-140 with foreground motions), ScanNet v2 [9] (100 indoor test videos sampled every third frame), and the full KITTI validation set [17] (13 outdoor driving videos with first 110 frames per sequence). This comprehensive benchmark covers synthetic animations, dynamic interactions, static indoor scenes, and real-world autonomous driving environments.

Quantitative results. We evaluate our method against strong baselines: DepthCrafter [23], known for video consistency, Metric3D [59], a state-of-the-art metric depth estimator, and ZoeDepth [1], a robust monocular depth estimation method. For fair comparison, we align all predictions within each video using a single shared scale factor across the entire scene with shift set to zero. As shown in

Frame Interval	Method	Sintel (∼50 frames)		Bonn (110 frames)		ScanNet (90 frames)		KITTI (110 frames)	
		AbsRel (↓)	$\delta 1$ (↑)	AbsRel (↓)	$\delta 1$ (↑)	AbsRel (↓)	$\delta 1$ (↑)	AbsRel (↓)	$\delta 1$ (↑)
1	DepthCrafter [23]	2.959	0.545	0.311	0.524	0.370	0.661	0.123	0.832
	Zoedepth [1]	0.454	0.356	0.277	0.562	0.347	0.436	0.216	0.633
	Metric3D [59]	0.372	0.610	**0.055**	**0.981**	0.059	0.974	**0.067**	**0.977**
	StableDepth (Ours)	**0.280**	**0.617**	0.063	0.979	**0.094**	**0.921**	0.112	0.900
4	DepthCrafter [23]	6.353	0.393	0.515	0.417	0.539	0.440	0.244	0.670
	Zoedepth [1]	0.461	0.357	0.277	0.563	0.349	0.436	0.218	0.629
	Metric3D [59]	0.361	**0.617**	**0.055**	**0.981**	0.059	0.974	**0.067**	**0.977**
	StableDepth (Ours)	**0.292**	0.606	0.063	0.979	**0.093**	**0.921**	0.112	0.901
8	DepthCrafter [23]	2.048	0.398	0.495	0.412	0.571	0.423	0.241	0.656
	Zoedepth [1]	0.470	0.357	0.276	0.562	0.352	0.434	0.218	0.633
	Metric3D [59]	0.363	**0.613**	**0.055**	**0.980**	0.060	0.973	**0.066**	**0.977**
	StableDepth (Ours)	**0.300**	0.604	0.062	**0.980**	**0.094**	**0.920**	0.112	0.902

Table 2. **Comparison of depth estimation methods across different frame intervals.** StableDepth maintains consistent performance across varying frame intervals while achieving superior accuracy. Best results are shown in **bold**, second best are underlined. Metric3D results on ScanNet are shown in gray as it was trained on this dataset.

Method	Bonn		ScanNet	
	AbsRel (↓)	δ_1 (↑)	AbsRel (↓)	δ_1 (↑)
Baseline	0.071	0.946	0.102	0.903
DPT	0.109	0.914	0.126	0.861
DPT + CNN	0.108	0.926	0.138	0.829
Dual DPT	**0.057**	**0.976**	**0.097**	**0.913**

Table 3. **Ablation on decoders.** The baseline model is trained only on labeled image data. We compare three approaches for handling unlabeled video data: using a single DPT decoder with reciprocal output (DPT), adding a lightweight CNN head (DPT + CNN), and using dual DPT decoders (Dual DPT).

Tab. 2, our method achieves superior performance across most datasets and metrics, with particularly strong improvements in AbsRel and δ_1 metrics compared to video-based approaches. We examine the impact of frame intervals on prediction accuracy to simulate varying temporal densities in real-world scenarios. Note that compared to our ablation studies, these results reflect extended training with additional iterations for optimal performance.

4.3. Ablation Study

Comparison of different decoders. In Tab. 3, we compare four methods using the Bonn and ScanNet v2 datasets: training with only labeled Hypersim data, and training with both labeled Hypersim [42] and unlabeled SA-V [41] data processed by three decoder designs. The first row (Baseline) shows results with only labeled Hypersim data, while the following rows show results with both labeled and unlabeled data. For labeled Hypersim data, depth in SCSI space is obtained directly from the DPT head output. For the unlabeled SA-V data, the method for obtaining disparity space depth varies: in the second row (DPT), it's the reciprocal of the original DPT output; in the third row (DPT + CNN), a lightweight CNN processes the DPT output; in the fourth row (Dual DPT), an additional DPT head provides the disparity space depth. The results show that using a single DPT head, either by reciprocating its output or adding a lightweight CNN, reduces prediction accuracy due to the partial inaccuracy of the pseudo-labels. However, using two independent DPT heads improves frame-to-frame consistency without impacting accuracy, validating the effectiveness of our disentangled approach.

Encoder feature stability. We analyze encoder feature stability across consecutive DAVIS frames to understand our dual-decoder effectiveness. Compared to DAv2, our method achieves 7.1% lower mean differences, 13.5% reduced standard deviation, and 2.3% higher cosine similarity. These improvements confirm that auxiliary decoder supervision enhances encoder stability, contributing to temporal consistency in depth predictions.

Effectiveness of unlabeled video data. As shown in Tab. 4, we evaluate incremental training data impact. Using only Hypersim [42] (row 1), then adding pseudo-labeled SA-V [41] data (row 2) shows mixed results - improving Bonn and ScanNet v2 but decreasing Sintel and KITTI performance. This reveals DepthCrafter's video consistency benefits but limitations in distant regions. Progressively incorporating VKITTI2 [6, 16], LightWheel [30], Hypersim [42], and MatrixCity [28] (rows 3-6) successfully addresses these limitations, achieving best performance across all datasets.

Comparison of video consistency. Tab. 5 showcases the performance comparison of depth estimation methods on the DAVIS dataset evaluated using Motion-aware temporal Difference (MTD (Eq. (9)), lower is better). The video diffusion-based DepthCrafter [23] achieves strong consistency (MTD=0.003479) but requires processing the entire video sequence and runs at only 0.45 FPS. While

Training Data	Sintel (∼50 frames)		Bonn (110 frames)		ScanNet (90 frames)		KITTI (110 frames)	
	AbsRel (↓)	$\delta 1$ (↑)	AbsRel (↓)	$\delta 1$ (↑)	AbsRel (↓)	$\delta 1$ (↑)	AbsRel (↓)	$\delta 1$ (↑)
Hypersim only [42]	**0.309**	0.578	0.071	0.946	0.102	0.903	0.122	0.871
+ SA-V [41]	0.326	0.561	**0.057**	0.976	0.097	0.913	0.143	0.823
+ VKITTI2 [6, 16]	0.318	0.577	0.058	0.976	0.097	0.911	0.129	0.839
+ LightWheel [30]	0.312	0.573	0.060	0.976	0.094	0.916	0.134	0.847
+ Hypersim-U [42]	0.314	0.572	0.064	0.976	0.096	0.918	0.123	0.870
+ Matrixcity [28]	0.315	**0.580**	0.062	**0.979**	**0.092**	**0.924**	**0.114**	**0.895**

Table 4. **Comparison of depth estimation performance with different training data configurations.** The baseline model is trained on labeled Hypersim data only, with subsequent rows showing the cumulative effect of adding different unlabeled datasets (VKITTI2: Virtual KITTI 2, Hypersim-U: treat Hypersim as unlabeled video data). Lower AbsRel (↓) and higher $\delta 1$ (↑) values indicate better performance. For each video segment, we align using only a single scale, setting the shift to zero.

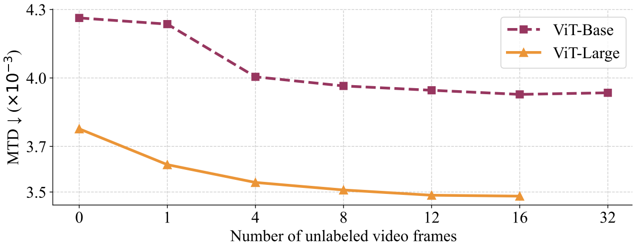

Figure 3. **Ablation on the number of unlabeled video frames in each batch.** From 0 to 16, more frames bring better scene-level consistency. (Out of memory for ViT-Large under 32 frames)

Method	MTD (↓)	Video Model?	Online?	FPS (↑)
DepthCrafter [23]	**0.003479**	✓	×	0.45
Baseline	0.003777	×	✓	**5.77**
StableDepth (Ours)	**0.003482**	×	✓	**5.77**

Table 5. **Performance comparison of different depth estimation methods on DAVIS dataset.** MTD: Motion-aware Temporal Difference (Eq. (9)). Online: whether the method can process data without requiring the entire video. FPS: inference speed in frames per second. StableDepth achieves comparable consistency to DepthCrafter that based on heavy SVD [3], while maintaining online capability and significantly faster inference.

our baseline (metric depth version of DAv2-Large) enables online inference at 5.77 FPS, it shows relatively higher inconsistency (MTD=0.003777). StableDepth significantly improves upon our baseline with a 7.8% reduction in MTD (0.003482), achieving comparable consistency to DepthCrafter. Crucially, StableDepth maintains online inference capability, processing frames independently without requiring past or future frames, and operates at 5.77 FPS on a H20 GPU - 13× faster than DepthCrafter. Both methods were evaluated on 1920×1080 resolution images using half-precision (FP16) for inference speed testing. These results demonstrate that StableDepth successfully combines strong temporal consistency with practical advantages in deployment, addressing a key limitation in current depth estimation approaches.

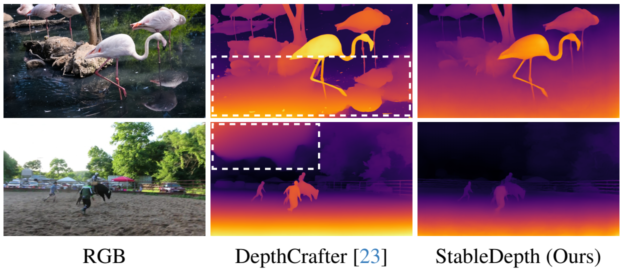

Figure 4. **Qualitative comparisons.** StableDepth shows improved robustness in handling ambiguous regions such as water surfaces and sky boundaries while maintaining accurate depth estimation for the main subjects (flamingo, horse) compared to DepthCrafter. The highlighted region in DepthCrafter results shows inconsistent depth predictions in these challenging regions.

Data	Pretrained	Bonn		ScanNet	
		AbsRel (↓)	δ_1 (↑)	AbsRel (↓)	δ_1 (↑)
L	DINOv2	0.086	0.929	0.098	0.911
L	DAv2	0.071	0.946	0.102	0.903
L + U	DINOv2	0.068	0.949	0.099	0.909
L + U	DAv2	**0.057**	**0.976**	**0.097**	**0.913**

Table 6. **Performance comparison on Bonn and ScanNet**. We compare different initialization strategies: DINOv2 pre-trained weights vs. DAv2's well-trained parameters. L: labeled data only; L+U: both labeled and unlabeled data with dual-decoder.

Comparison of unlabeled video frames. As shown in Fig. 3, MTD scores improve steadily with more unlabeled video frames in each mini-batch and gradually converge at 16 frames, with negligible gains beyond this point, supporting our choice of 16-frame clips as an optimal trade-off.

Comparison of different pre-trained encoders. Tab. 6 shows the effectiveness of our semi-supervised training strategy on the Bonn and ScanNet datasets. We compare two initialization strategies: DINOv2 and DAv2 pre-trained

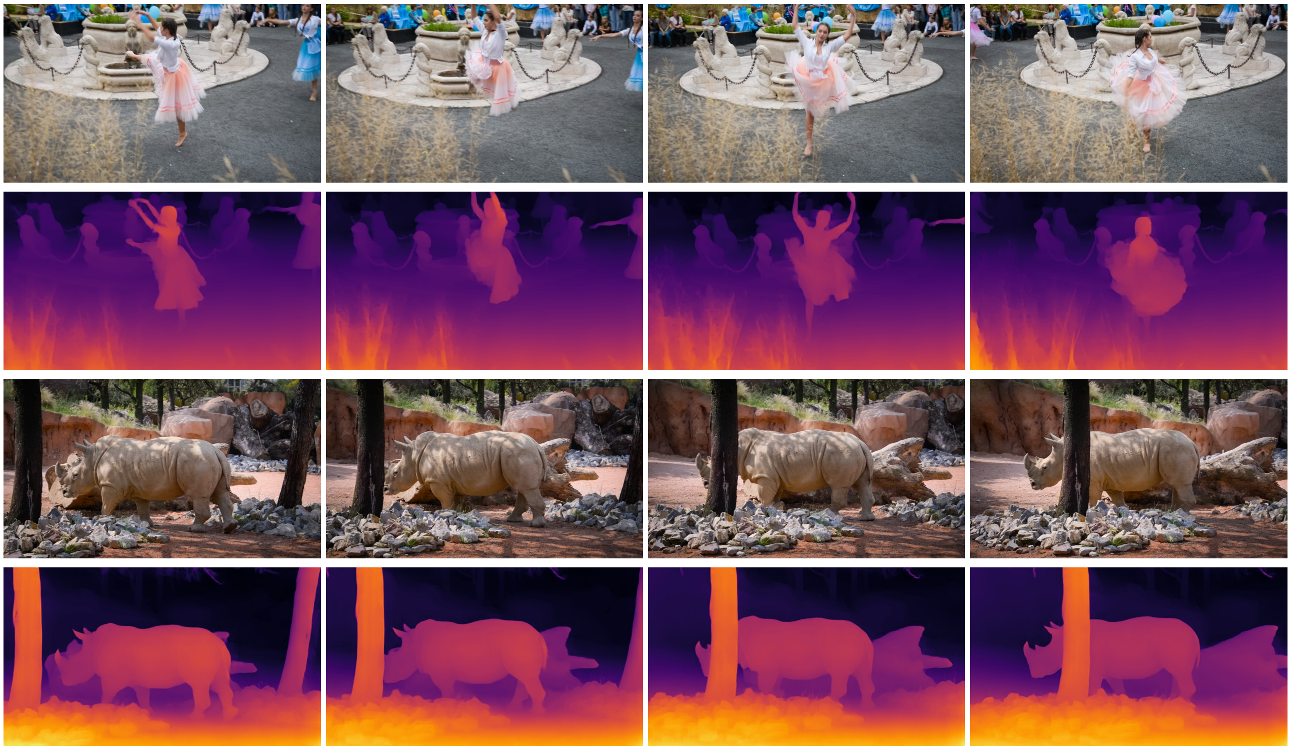

Figure 5. **Qualitative visualizations of depth predictions across diverse scenarios.** Each row pair shows an input sequence and corresponding depth predictions from our method. Our approach maintains consistent depth estimation while preserving fine geometric details across challenging cases including dynamic motions (dancer), reflective surfaces (water), and large objects (rhinoceros).

weights. With only labeled data (L), DAv2 initialization outperforms DINOv2, particularly on Bonn (0.071 vs 0.086 AbsRel). Incorporating unlabeled video data (L+U) through our dual-decoder approach yields consistent improvements across both datasets. DAv2 with L+U achieves the best performance, reducing AbsRel to 0.057 on Bonn and 0.097 on ScanNet, demonstrating the effectiveness of our semi-supervised approach.

4.4. Qualitative Results

Our method demonstrates robust performance across diverse real-world scenarios. As shown in Fig. 1, it generates stable depth maps even in challenging conditions, capturing dynamic motions in dog agility courses, street performances, and natural scenes while maintaining consistent geometry frame-by-frame. Fig. 4 shows our method's superiority over DepthCrafter in ambiguous depth cue cases. In the flamingo scene, DepthCrafter produces inconsistent depth predictions for water reflections, while our method maintains coherent depth across the water surface. In the horse sequence, DepthCrafter mistakenly estimates the sky as foreground, whereas our approach handles the background-sky transition more accurately. These results highlight our model's robustness to common depth estimation challenges like reflective surfaces and infinite-depth regions, while preserving scene consistency. Fig. 5 further demonstrates the effectiveness of our method in challenging scenarios, including fast-moving subjects (dancer) and large objects (rhinoceros). Our approach consistently preserves fine geometric details and maintains stable depth relationships across frames, even under complex lighting conditions and dynamic motions. The temporal smoothness achieved enables reliable depth estimation for practical applications requiring consistent 3D understanding.

5. Conclusion

We present StableDepth, a framework that bridges the gap between relative and metric depth estimation through our scene-consistent and scale-invariant (SCSI) paradigm. By leveraging our dual-decoder architecture and video diffusion baking strategy, we effectively decouple depth accuracy from temporal consistency learning, while harnessing knowledge from unlabeled video data. StableDepth enables online inference at 13× faster speed than previous video-based methods while maintaining comparable temporal stability. This approach advances monocular depth estimation toward more practical applications in 3D reconstruction, robotics, and augmented reality.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 62422606, 62201484) and DAMO Academy Research Intern Program.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv:2302.12288*, 2023. 5, 6

[2] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv:2307.14460*, 2023. 4

[3] 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:2311.15127*, 2023. 2, 5, 7

[4] 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. In *ICLR*, 2025. 2, 3

[5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 2, 5

[6] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv:2001.10773*, 2020. 5, 6, 7

[7] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *CVPR*, 2025. 3

[8] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *NeurIPS*, 2016. 3

[9] 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, 5

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4, 5

[11] 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 *ICCV*, 2021. 3

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 1, 3

[13] Fatima El Jamiy and Ronald Marsh. Survey on depth perception in head mounted displays: distance estimation in virtual reality, augmented reality, and mixed reality. *IET Image Processing*, 2019. 2

[14] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *SCIA*, 2003. 5

[15] 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 *ECCV*, 2024. 2

[16] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 5, 6, 7

[17] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 2, 5

[18] Ming Gui, Johannes S Fischer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. In *AAAI*, 2025. 2

[19] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *ICCV*, 2023. 3

[20] S Mahdi H. Miangoleh, Mahesh Reddy, and Yağız Aksoy. Scale-invariant monocular depth estimation via ssi depth. In *SIGGRAPH*, 2024. 2

[21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 3

[22] 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. *TPAMI*, 2024. 2, 3

[23] 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. In *CVPR*, 2025. 2, 3, 4, 5, 6, 7

[24] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *CVPR*, 2022. 3

[25] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 2

[26] Bingxin Ke, Dominik Narnhofer, Shengyu Huang, Lei Ke, Torben Peters, Katerina Fragkiadaki, Anton Obukhov, and Konrad Schindler. Video depth without video models. In *CVPR*, 2025. 3

[27] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *CVPR*, 2021. 3

[28] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *ICCV*, 2023. 5, 6, 7

[29] 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 *WACV*, 2023. 3

[30] LightwheelAI and LightwheelOcc contributors. Lightwheelocc: A 3d occupancy synthetic dataset in autonomous driving. https://github.com/OpenDriveLab/LightwheelOcc, 2024. 5, 6, 7

[31] Nian Liu, Ni Zhang, Ling Shao, and Junwei Han. Learning selective mutual attention and contrast for rgb-d saliency detection. *TPAMI*, 2021. 3

[32] Xingtong Liu, Ayushi Sinha, Masaru Ishii, Gregory D Hager, Austin Reiter, Russell H Taylor, and Mathias Unberath. Dense depth estimation in monocular endoscopy with self-supervised learning methods. *TMI*, 2019. 2

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[34] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. In *SIGGRAPH*, 2020. 3

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023. 2

[36] E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. In *IROS*, 2019. 2, 5

[37] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, 2024. 2, 3

[38] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv:2502.20110*, 2025. 3

[39] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 2, 3, 4

[40] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 3, 4, 5

[41] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, 2025. 5, 6, 7

[42] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 3, 5, 6, 7

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[44] Abhinav Sagar. Monocular depth estimation using multi scale neural network and feature fusion. In *WACV*, 2022. 3

[45] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *TPAMI*, 2008. 1

[46] Mohamed Sayed, John Gibson, Jamie Watson, Victor Prisacariu, Michael Firman, and Clément Godard. Simplerecon: 3d reconstruction without 3d convolutions. In *ECCV*, 2022. 3

[47] Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors. 2025. 2, 3

[48] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *ECCV*, 2020. 2

[49] Chaoyang Wang, Simon Lucey, Federico Perazzi, and Oliver Wang. Web stereo video supervision for depth prediction from dynamic scenes. In *3DV*, 2019. 3

[50] Lijun Wang, Yifan Wang, Linzhao Wang, Yunlong Zhan, Ying Wang, and Huchuan Lu. Can scale-consistent monocular depth be learned in a self-supervised scale-invariant manner? In *ICCV*, 2021. 3

[51] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, 2019. 2

[52] Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *ICCV*, 2023. 3

[53] Jamie Watson, Mohamed Sayed, Zawar Qureshi, Gabriel J Brostow, Sara Vicente, Oisin Mac Aodha, and Michael Firman. Virtual occlusions through implicit depth. In *CVPR*, 2023. 3

[54] Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *ICRA*, 2019. 2

[55] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *CVPR*, 2020. 3

[56] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 2, 3, 4

[57] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, 2024. 3, 4, 5

[58] Rajeev Yasarla, Hong Cai, Jisoo Jeong, Yunxiao Shi, Risheek Garrepalli, and Fatih Porikli. Mamo: Leveraging memory and attention for monocular video depth estimation. In *ICCV*, 2023. 3

[59] 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 *ICCV*, 2023. 2, 3, 5, 6

[60] Sen Zhang, Jing Zhang, and Dacheng Tao. Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating imu motion dynamics. In *ECCV*, 2022. 3

[61] Ruijie Zhu, Chuxin Wang, Ziyang Song, Li Liu, Tianzhu Zhang, and Yongdong Zhang. Scaledepth: Decomposing metric depth estimation into scale prediction and relative depth estimation. *arXiv:2407.08187*, 2024. 3