FLEXDEPTH++: TOWARDS A UNIFIED SELF-SUPERVISED MULTI-FRAME MONOCULAR DEPTH ESTIMATION FRAMEWORK FOR DRIVING AND LOW-ALTITUDE AERIAL SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous driving and low-altitude UAV systems both demand accurate self-supervised depth estimation by exploiting multiple frames. However, building a unified framework across these two domains remains challenging, as their substantial differences in depth distributions and camera motion lead to inaccurate and unstable temporal matching. To address this challenge, we propose *FlexDepth++*, a unified framework for driving and low-altitude aerial scenes based on hierarchical cost-volume enhancement. Specifically, Multi-resolution Cost-Volume Construction aggregates complementary matching cues from different feature resolutions to enrich geometric context, while Stability-guided Cost-Volume Refinement exploits prediction discrepancies across training stages to derive a stability prior for refining unreliable cost-volume responses. Together, the two modules follow a construction-then-refinement paradigm to improve both the expressiveness of temporal geometric matching across heterogeneous scenarios. Extensive experiments on UAVScenes, UAVLight, KITTI, and Cityscapes show that FlexDepth++ achieves state-of-the-art performance on all four benchmarks, including an Abs Rel of **0.071** on UAVScenes and **0.082** on KITTI. To the best of our knowledge, FlexDepth++ is among the first unified self-supervised multi-frame monocular depth estimation frameworks systematically validated across both autonomous driving and low-altitude aerial scenes. Code will be released.

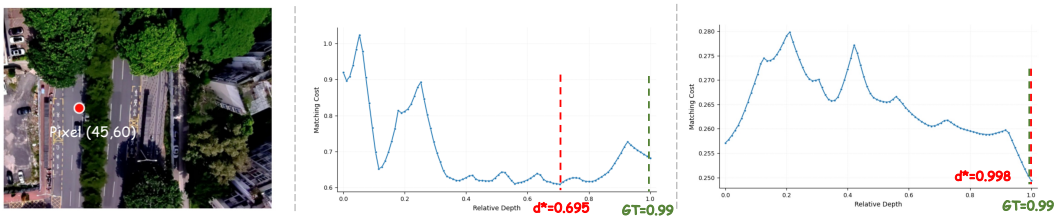
1 INTRODUCTION

The rapid development of autonomous driving and low-altitude aerial systems, particularly unmanned aerial vehicles (UAVs), has created a growing demand for robust 3D perception across heterogeneous camera platforms Yurtsever et al. (2020). Both domains critically rely on dense geometric understanding to support applications such as object detection and autonomous navigation, even though they differ substantially in camera motion, depth distribution, and scene layout Li et al. (2026); Wu et al. (2025b). Monocular depth estimation, as a fundamental means of recovering dense scene geometry from visual observations, has therefore become an important perception capability for both driving and low-altitude aerial platforms Zhao et al. (2022); Zhou et al. (2025).

However, scaling monocular depth estimation across such diverse real-world environments remains challenging, as collecting large-scale and high-quality depth annotations is expensive and labor-intensive. Self-supervised monocular depth estimation provides a practical alternative by learning scene geometry directly from unlabeled videos Wu et al. (2025a); Gao et al. (2026). Building on this paradigm, multi-frame methods further exploit the temporal geometric information naturally contained in video sequences to reduce monocular depth ambiguity through cross-frame consistency Watson et al. (2021); Woo et al. (2024); Zhou et al. (2025). These shared perception demands and learning principles naturally call for a *unified self-supervised multi-frame monocular depth estimation framework that can operate across both driving and low-altitude aerial scenes*.

However, achieving such a unified depth estimation framework is non-trivial, as autonomous driving and low-altitude aerial scenarios exhibit markedly different scene geometry and camera motion

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Query Pixel in Low-Altitude Aerial Scene

|d* - GT| = 0.295

(b) Conventional Cost-Volume Profile

|d*. - GT| = 0.008

(c) FlexDepth++ Cost-Volume Profile

Figure 1: Comparison of cost-volume matching profiles in a low-altitude aerial scene. (a) A representative query pixel. (b) Conventional cost-volume matching produces an ambiguous profile whose globally optimal depth hypothesis deviates substantially from the ground truth. (c) FlexDepth++ produces a more reliable matching profile, with the global minimum closely aligned with the ground-truth depth. Here, $d^* = \arg\min_d V(d)$ denotes the depth hypothesis corresponding to the minimum matching cost, where d and matching cost represent the horizontal and vertical axes, respectively. For visualization, d are normalized to relative depth.

patterns. Compared with conventional driving environments, low-altitude aerial scenes typically involve broader depth ranges, more flexible camera motion, and rapid viewpoint changes, making stable cross-frame correspondence considerably more difficult Madhuanand et al. (2021); Song et al. (2026); Ye et al. (2026). Such domain disparities pose particular challenges to multi-frame methods, which commonly realize temporal geometric matching through cost volumes by comparing neighboring-frame features across multiple candidate depth hypotheses. Specifically, conventional cost-volume designs encounter two key limitations under these heterogeneous scenarios. The first is *Insufficient Matching Context*: cost volumes constructed at a single feature resolution provide limited complementary texture and structural information, making it difficult to simultaneously capture broad geometric context and fine-grained local correspondence. The second is *Temporal Matching Instability*: complex camera motion and ambiguous visual correspondence can produce unstable and less discriminative matching responses, whose uncertainty is further reflected by inconsistent depth predictions during training. As illustrated in Figure 1(b), directly applying driving-oriented cost-volume designs to low-altitude aerial videos often results in ambiguous matching profiles, characterized by spurious local minima and flattened response regions, making reliable depth hypotheses harder to distinguish.

To this end, we propose **FlexDepth++**, a unified self-supervised multi-frame monocular depth estimation framework for both driving and low-altitude aerial scenes, built upon a hierarchical cost-volume enhancement strategy. To address the two limitations above, we first introduce Multi-resolution Cost-Volume Construction (MCVC), which constructs cost volumes from complementary feature resolutions and further aggregates them to enrich temporal matching context with both broad geometric structure and fine-grained local details. Moreover, we develop Stability-guided Cost-Volume Refinement (SCVR), which derives a stability mask from prediction discrepancies across training stages and uses it as an auxiliary spatial prior to refine cost-volume representations toward more stable geometric regions. Together, these two components follow a construction-then-refinement paradigm to progressively enhance cost-volume representations, providing richer and more robust temporal geometric cues for depth prediction across heterogeneous driving and low-altitude aerial scenarios.

We conduct extensive experiments on UAVScene Wang et al. (2025c), UAVLight Du et al. (2026a), KITTI Geiger et al. (2013), and Cityscapes Cordts et al. (2016), We conduct extensive experiments on two low-altitude aerial datasets, UAVScene and UAVLight, and two autonomous driving datasets, KITTI and Cityscapes. FlexDepth++ achieves state-of-the-art performance across all four benchmarks, demonstrating strong effectiveness and generalization across heterogeneous driving and aerial scenarios. In particular, it achieves an Abs Rel of **0.071** on UAVScenes and **0.082** on KITTI, highlighting its strong performance on both low-altitude aerial and autonomous driving scenes. To the best of our knowledge, FlexDepth++ is among the first unified self-supervised multi-frame monocular depth estimation frameworks systematically designed and validated across both domains.

Our main contributions are summarized as follows:

- We propose FlexDepth++, a unified self-supervised multi-frame monocular depth estimation framework for both driving and low-altitude aerial scenes. To the best of our knowledge, it is among the first frameworks systematically designed to operate across these two heterogeneous domains within a unified setting.
- We introduce a hierarchical cost-volume enhancement paradigm that follows a construction-then-refinement strategy, combining Multi-resolution Cost-Volume Construction and Stability-guided Cost-Volume Refinement to enrich temporal matching context and improve the robustness of cross-frame geometric cues.
- Extensive experiments on UAVScenes, UAVLight, KITTI, and Cityscapes demonstrate state-of-the-art performance across both low-altitude aerial and autonomous driving domains, achieving an Abs Rel of 0.071 on UAVScenes and 0.082 on KITTI.

2 Related Work

2.1 Self-supervised Monocular Depth Estimation in Driving Scenes

Self-supervised monocular depth estimation learns depth from unlabeled monocular videos by exploiting photometric consistency across adjacent frames Zhou et al. (2017); Godard et al. (2019); Gao et al. (2026). A depth network predicts the target-view depth, while a pose network estimates relative camera motion for differentiable view synthesis, avoiding the need for dense depth annotations. Early works established this joint depth-and-pose learning paradigm Zhou et al. (2017), while Monodepth2 Godard et al. (2019) further improved training robustness through minimum reprojection and auto-masking. These methods form the standard self-supervised monocular depth estimation pipeline and have achieved strong performance on driving benchmarks such as KITTI and Cityscapes Geiger et al. (2013); Cordts et al. (2016).

To further exploit temporal information, recent methods extend single-frame depth prediction to multi-frame formulations Watson et al. (2021); Casser et al. (2019); Han et al. (2026). Rather than relying solely on monocular appearance cues, they incorporate consecutive frames and explicitly exploit cross-frame geometric correspondence. Among them, ManyDepth Watson et al. (2021) constructs a cost volume from temporally adjacent frames and combines multi-frame matching cues with monocular priors, establishing a representative cost-volume-based paradigm for self-supervised depth estimation. Subsequent works further improve temporal consistency, matching quality, and pose robustness, but are predominantly developed and evaluated in structured driving environments.

2.2 Self-supervised Monocular Depth Estimation in UAV Scenes

Existing UAV vision research has primarily focused on 2D perception tasks such as object detection, tracking, active search, and visual localization Huang et al. (2026); Ye et al. (2026); Xue et al. (2026); Qin et al. (2026). More recently, aerial 3D perception and depth estimation have attracted increasing attention. Datasets such as UAVScenes Wang et al. (2025c) and UAVLight Du et al. (2026a), together with benchmarks such as AerialMetric Song et al. (2026), reveal the challenges of aerial depth perception caused by viewpoint changes, altitude variations, complex camera motion, and the domain gap between ground-level and aerial imagery.

Existing monocular depth methods for UAV scenes mainly rely on either zero-shot depth foundation models, such as Depth Anything V2 Yang et al. (2024b), or adapted open-domain geometry models, such as MoGe Wang et al. (2025a) and MoGe-2 Wang et al. (2025b). While these methods benefit from strong monocular geometry priors, they are predominantly designed for single-image depth prediction and do not explicitly exploit temporal geometric constraints in UAV videos. Consequently, self-supervised multi-frame monocular depth estimation for low-altitude aerial scenes remains largely underexplored, especially under broader depth distributions and complex camera motion.

3 Methodology

As depicted in Figure 2, the pipeline consists of four main components. In Section 3.1, we first introduce Feature Extraction and Geometric Warping, which extracts multi-level target and source fea-

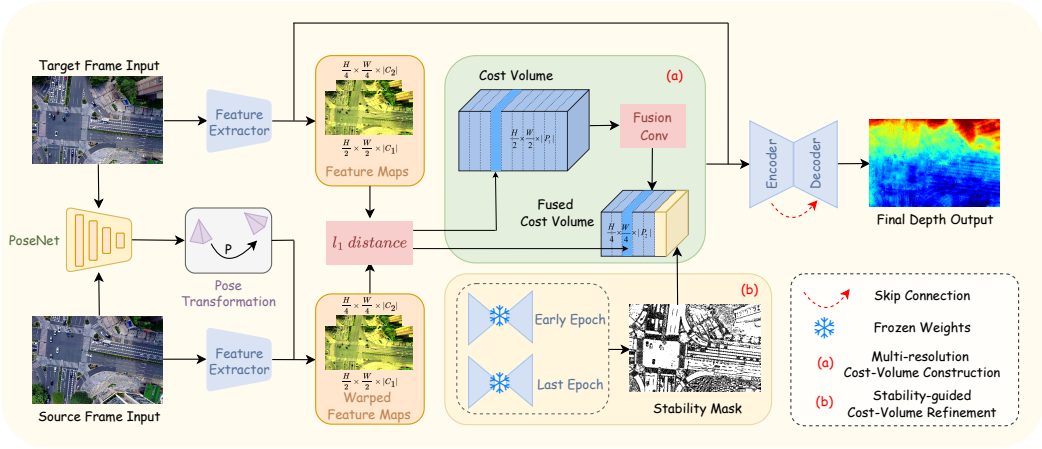


Figure 2: Overview of the proposed FlexDepth++. Given target and source frames, multi-level features are extracted, while the source features are simultaneously warped to the target view using the estimated relative pose. Initial cost volumes are then constructed via L_1 matching over candidate depth hypotheses. Our hierarchical cost-volume enhancement strategy then consists of: (a) Multi-resolution Cost-Volume Construction (MCVC), which aggregates complementary cost volumes from different feature resolutions; and (b) Stability-guided Cost-Volume Refinement (SCVR), which derives a stability mask from prediction discrepancies across training stages to refine the cost-volume representation. Finally, the enhanced cost volumes are combined with image features for dense depth prediction.

tures and geometrically aligns the source features to prepare for subsequent cost-volume construction. Next, Sections 3.2 and 3.3 present Multi-resolution Cost-Volume Construction (MCVC) and Stability-guided Cost-Volume Refinement (SCVR), respectively, which progressively enhance temporal matching representations. Furthermore, Section 3.4 introduces the Depth Prediction Network, which integrates the enhanced cost-volume representations with image features through an encoder-decoder architecture to produce the final dense depth prediction. Finally, Section 3.5 presents the self-supervised training objective.

3.1 Feature Extraction and Warping

Given a target frame I_t and its neighboring source frames I_s, FlexDepth++ aims to predict the target-view depth $\hat{D}_t$ in a self-supervised manner. Following the standard multi-frame self-supervised depth estimation paradigm Watson et al. (2021), we first extract multi-level image features F_t^l and F_s^l from the target and source frames, respectively, where l denotes the feature level. Meanwhile, a pose network estimates the relative camera transformation between the target and source views, denoted as $T_{t\to s} \in SE(3)$, which encodes the relative rotation and translation from the target frame to the source frame. To establish cross-frame geometric correspondence, the source features are warped toward the target view under a set of candidate depth hypotheses $\mathcal{D} = \{d_k\}_{k=1}^{D}$. For each candidate depth d_k, the warped source feature is obtained as

$$\hat{F}_{s\to t}^{l,k} = \mathcal{W}\left(F_s^l, T_{t\to s}, K, d_k\right), \tag{1}$$

where K denotes the camera intrinsic matrix and $\mathcal{W}(\cdot)$ represents the differentiable geometric warping operation. The resulting target features F_t^l and warped source features $\hat{F}_{s\to t}^{l,k}$ provide geometrically aligned feature pairs under different depth hypotheses, which serve as the basis for the subsequent cost-volume construction.

3.2 Multi-resolution Cost-Volume Construction

To enrich temporal matching context across heterogeneous depth distributions and scene structures, we construct cost volumes at multiple feature resolutions and aggregate their complementary geometric cues. Specifically, given the target feature F_t^l and the warped source feature $\hat{F}_{s\to t}^{l,k}$ at feature

level l and depth hypothesis d_k, we compute their feature-wise matching cost using the L_1 distance:

$$V^{l,k}(u,v) = \frac{1}{C_l} \left\| F_t^l(u,v) - \hat{F}_{s\to t}^{l,k}(u,v) \right\|_1, \tag{2}$$

where C_l denotes the feature dimension at level l. By evaluating the matching cost over all candidate depth hypotheses $\mathcal{D} = \{d_k\}_{k=1}^{D}$, we obtain the cost volume corresponding to the $1/4$-resolution feature map:$V^{1/4} \in \mathbb{R}^{B\times D\times H_{1/4}\times W_{1/4}}$. Simultaneously, we construct another cost volume $V^{1/2}$ from the $1/2$-resolution feature maps using the same matching procedure. The higher-resolution representation preserves finer textures and structural details, enabling the L_1 matching distance to capture more precise cross-frame correspondence around object boundaries and high-frequency regions.

To further integrate these multi-resolution matching representations, we introduce a lightweight *Fusion Conv* module. Specifically, higher-resolution cost volumes are first downsampled to a common spatial resolution and then processed by a stack of 3×3 convolutional layers, which simultaneously reduce the channel dimension and aggregate complementary matching cues across resolutions. The resulting fused cost volume can be written as

$$V_{\text{fuse}} = \Phi_{\text{fuse}} \left(\text{Concat} \left[V^{l_1}, \mathcal{D}(V^{l_2}), \ldots \right] \right), \tag{3}$$

where $\mathcal{D}(\cdot)$ denotes the downsampling operation and $\Phi_{\text{fuse}}(\cdot)$ represents the stacked 3×3 convolutions for channel reduction and multi-resolution feature aggregation.

3.3 Stability-Guided Cost-Volume Refinement

Although multi-resolution cost-volume construction enriches temporal matching context, the resulting matching responses may still be unreliable in regions affected by complex camera motion, weak textures, or ambiguous correspondence. To further improve the robustness of temporal geometric representations, we propose Stability-guided Cost-Volume Refinement (SCVR), which exploits prediction stability across training stages as an auxiliary spatial prior for cost-volume refinement.

Specifically, during the second training stage, we compute two disparity maps using model weights saved from the initial and final epochs of the first stage, denoted as disp_i and disp_l, respectively. Inspired byZhang et al. (2025), the discrepancy between these two disparity maps is used to characterize the stability of depth prediction during training: regions with smaller discrepancies generally correspond to more stable geometric predictions, whereas regions with large variations are more likely to correspond to uncertain or ambiguous areas. Based on this observation, we construct a binary stability mask M by comparing the disparity discrepancy against an adaptive threshold:

$$M = \left[\left| \text{disp}_i^{\beta} - \text{disp}_l^{\beta} \right| < g\left(\phi(f_{s_i}, f_{s_l})\right) \right], \tag{4}$$

where $[\cdot]$ denotes the Iverson bracket, β is a scaling factor, and $g(\phi(f_{s_i}, f_{s_l}))$ serves as an adaptive thresholding function.

The resulting mask M therefore encodes spatial prediction stability obtained from the first-stage training dynamics. Instead of directly imposing handcrafted weights on the cost volume, we use M as an auxiliary spatial prior to augment the fused cost-volume representation. This allows the subsequent depth prediction network to distinguish relatively stable geometric regions from uncertain ones and learn to refine temporal matching cues in a data-driven manner.

3.4 Depth Prediction Network

Encoder. We adopt a standard multi-scale feature extraction architecture based on YOLOv11 Khanam & Hussain (2024) as the depth encoder. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the encoder progressively extracts a hierarchy of feature maps through successive downsampling operations:

$$\{\mathbf{F}_{enc,1}, \mathbf{F}_{enc,2}, \ldots, \mathbf{F}_{enc,L}\} = E(I), \tag{5}$$

where $\mathbf{F}_{enc,i}$ denotes the encoder feature at the i-th scale. These hierarchical features capture visual representations at different spatial resolutions and provide appearance cues and skip connections for subsequent depth decoding.

Decoder. Following FlexDepth Zhu et al. (2026), we employ an efficient upsampling-centric depth decoder. At each decoding stage, the feature from the previous stage is first upsampled and concatenated with the corresponding encoder feature through a skip connection, followed by a post-convolution module and a depth prediction head. In FlexDepth++, we further inject the enhanced cost-volume representation into the corresponding decoding stages. For the scales where cost-volume features are available, the decoding process is formulated as

$$D_i = \mathcal{P}_i \left\{ \mathcal{C}_i \left(\text{Concat} \left[\mathcal{U}_i(\mathbf{F}_{i+1}), \mathbf{F}_{enc,i}, \widetilde{V}_i \right] \right) \right\}, \tag{6}$$

where $\mathcal{U}_i(\cdot)$, $\mathcal{C}_i(\cdot)$, and $\mathcal{P}_i(\cdot)$ denote the upsampling operation, post-convolution module, and depth prediction head, respectively. $\widetilde{V}_i$ denotes the cost-volume representation enhanced by the proposed MCVC and SCVR modules. For scales without cost-volume features, the original FlexDepth decoding operation is retained. The decoder progressively recovers spatial details and ultimately produces the final dense depth prediction $\hat{D}_t$.

3.5 Self-Supervised Training Objective

We train FlexDepth++ using the standard self-supervised view-reconstruction objective following Godard et al. (2019). The predicted depth and relative camera pose are used to synthesize the target view from neighboring source images, and supervision is provided by a photometric reprojection loss combining SSIM and L_1 terms, together with minimum reprojection, auto-masking, and edge-aware smoothness regularization. For multi-frame training, we further follow Watson et al. (2021) and employ a consistency constraint between the single-frame depth prediction and the depth hypothesis indicated by the cost volume. Regions with large disagreement are treated as unreliable temporal matches and receive additional consistency supervision. The final training objective combines photometric reconstruction, multi-frame consistency, and edge-aware smoothness losses.

4 Experiments

4.1 Datasets and Metrics

Datasets. We evaluate FlexDepth++ on two low-altitude UAV datasets and two autonomous driving datasets. UAVScene Wang et al. (2025c) contains four large outdoor scenes: AMtown, AMvalley, HKairport, and HKisland. These scenes cover diverse low-altitude UAV environments, including urban areas, valley-like natural regions, airport scenes. With synchronized image-LiDAR data, reconstructed 3D maps, and accurate 6-DoF poses, UAVScene provides a challenging benchmark for evaluating self-supervised depth estimation under complex UAV motion and broad depth distributions. UAVLight Du et al. (2026a) are captured along repeatable, geo-referenced flight paths at multiple fixed times of day, resulting in consistent geometry but different lighting conditions, such as sunlight direction, exposure changes. KITTI Geiger et al. (2013) contains vehicle-mounted driving sequences with structured road scenes and relatively regular forward motion, and is used as the standard benchmark for self-supervised monocular depth estimation. Cityscapes Cordts et al. (2016) contains urban street-view videos with diverse city layouts, vehicles, and is used to evaluate cross-dataset generalization in driving scenes.

Metrics. Following standard evaluation protocols for self-supervised monocular depth estimation, we report seven commonly used metrics: Abs Rel, Sq Rel, RMSE, RMSE log, and the threshold accuracies δ_1, δ_2, and δ_3. For error metrics, including Abs Rel, Sq Rel, RMSE, and RMSE log, lower values indicate better performance. For threshold accuracies, higher values are better, where δ_k measures the percentage of pixels satisfying $\max(\hat{D}/D, D/\hat{D}) < 1.25^k$, with $k \in \{1, 2, 3\}$.

4.2 Implement Details

FlexDepth++ is trained on monocular triplets $\{t\text{–}1, t, t\text{+}1\}$. The multi-frame stage is initialized from the FlexDepth Zhu et al. (2026) single-frame checkpoint of the corresponding dataset, which also serves as the frozen consistency teacher. KITTI and Cityscapes adopt the X-scale model, whereas UAVScenes and UAVLight adopt the more compact M-scale model. We optimize with NAdam at a learning rate of 5×10^{-5} on KITTI, 8×10^{-5} on Cityscapes, and 1×10^{-4} on both UAV datasets, with

Table 1: Quantitative comparison on UAVScenes Wang et al. (2025c) and UAVLight Du et al. (2026a). Input resolutions are 384×320 and 640×480, respectively. **M**: monocular videos; **M+F**: monocular videos with optical flow; **Pre.**: pretrained depth model; **T.F.**: test input frames; *: zero-shot evaluation. **Bold** denotes the best results.

Method	Year	T.F.	Data	Depth Error↓				Depth Accuracy↑		
				Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
UAVScenes										
Depth Anything V2 (ViT-S)* [Yang et al.]	2024	1	Pre.	0.184	6.890	22.410	0.227	0.731	0.917	0.978
Depth Anything V2 (ViT-L)* [Yang et al.]	2024	1	Pre.	0.247	10.680	28.860	0.300	0.585	0.839	0.953
ManyDepth2 [Zhou et al.]	2025	$2, (-1, 0)$	M+F	0.079	0.869	7.325	0.100	0.942	0.998	**1.000**
FlexDepth[Zhu et al.]	2026	1	M	0.081	0.930	7.524	0.102	0.935	0.998	**1.000**
FlexDepth++		$2, (-1, 0)$	M	**0.071**	**0.722**	**6.744**	**0.092**	**0.954**	**0.999**	**1.000**
UAVLight										
ManyDepth [Watson et al.]	2021	$2, (-1, 0)$	M	0.074	0.925	8.527	0.092	0.973	0.996	0.999
Depth Anything V2 (ViT-S)* [Yang et al.]	2024	1	Pre.	0.333	17.530	39.510	0.413	0.426	0.701	0.890
Depth Anything V2 (ViT-L)* [Yang et al.]	2024	1	Pre.	0.381	22.270	44.810	0.465	0.369	0.632	0.842
ManyDepth2 [Zhou et al.]	2025	$2, (-1, 0)$	M+F	0.069	0.843	8.016	0.087	0.977	0.996	0.999
FlexDepth[Zhu et al.]	2026	1	M	0.048	0.431	5.838	0.063	0.990	0.998	**1.000**
FlexDepth++		$2, (-1, 0)$	M	**0.046**	**0.380**	**5.550**	**0.060**	**0.992**	**0.999**	**1.000**

effective batch size 12. The multi-frame stage runs 30 epochs on all datasets with a step scheduler (step size 15, decay 0.1). The training depth range is $[0.1, 100]$ m for driving scenes and $[0.1, 200]$ m for UAV scenes, and the cost volume uses 96 depth bins as in ManyDepth Watson et al. (2021). The consistency weight is 0.1 on KITTI and UAVScenes and 1.0 on Cityscapes and UAVLight, as the latter two contain more complex scenes with dynamic objects and illumination changes.

4.3 EXPERIMENTAL RESULTS

Main Results of UAV Scenes. As shown in Table 1, FlexDepth++ consistently establishes new state-of-the-art performance across both low-altitude aerial and autonomous driving benchmarks. On UAVScenes, FlexDepth++ reduces Abs Rel from 0.081 to 0.071 over FlexDepth Zhu et al. (2026), together with clear improvements in Sq Rel, RMSE, and $\delta < 1.25$, demonstrating the benefit of enhanced multi-frame geometric reasoning in challenging aerial scenes. Compared with zero-shot pretrained depth models Depth Anything v2 Yang et al. (2024b) FlexDepth++ also exhibits a substantial advantage on UAVScenes, suggesting that generic monocular priors alone are insufficient to bridge the pronounced domain gap in low-altitude aerial scenes, whereas self-supervised temporal geometry provides effective scene-adaptive cues. A similar trend is observed on UAVLight, where FlexDepth++ achieves an Abs Rel of 0.046 and RMSE of 5.550, outperforming both the single-frame FlexDepth baseline and the classical multi-frame ManyDepth Watson et al. (2021).

Main Results of Driving Scenes. Notably, the improvements are not limited to aerial scenarios. As shown in Table 2, the canonical benchmark for self-supervised monocular depth estimation, FlexDepth++ establishes a new state of the art with an Abs Rel of 0.082 and an RMSE of 3.901. It surpasses representative multi-frame methods including ManyDepth(0.098) Watson et al. (2021), ProDepth (0.086) Woo et al. (2024), ManyDepth2 (0.091) Zhou et al. (2025), and VFM-Dynamo (0.086) Du et al. (2026b), while also improving over the FlexDepth baseline by a clear margin. On Cityscapes, FlexDepth++ further achieves the best Abs Rel of 0.083 and RMSE of 5.145, outperforming recent methods such as DSI-MonoViT Zhang et al. (2025)and ProDepth Woo et al. (2024). These results demonstrate that FlexDepth++ improves depth estimation in challenging low-altitude aerial scenes while simultaneously advancing performance on conventional driving benchmarks, supporting the effectiveness of the proposed hierarchical cost-volume enhancement across markedly different camera motions, depth distributions, and scene structures.

Effect of each Module in FlexDepth++. To evaluate the contribution of each proposed component, we conduct ablation studies on both KITTI and UAVScenes, and report the results in Table 2. Method **(a)** is the baseline model without MCVC or SCVR, which achieves an Abs Rel of 0.084 on KITTI and serves as the reference setting for subsequent comparisons. Method **(b)** extends **(a)** by introducing the Multi-resolution Cost-Volume Construction (MCVC) module. On KITTI, MCVC improves Abs Rel from 0.084 to 0.082, reduces Sq Rel from 0.554 to 0.545, and increases $\delta < 1.25$ from 0.923 to 0.926. On UAVScenes, the gain is more pronounced, yielding an Abs Rel

Table 2: Quantitative comparison on KITTI (Eigen split) Geiger et al. (2013) and Cityscapes Cordts et al. (2016). Input resolutions are 640×192 and 416×128, respectively. **M**: monocular videos; **M+F**: monocular videos with optical flow; **M+Se**: monocular videos with semantic segmentation; **T.F.**: test input frames. **Bold** denotes the best results.

Method	Year	T.F.	Data	Depth Error↓				Depth Accuracy↑		
				Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
KITTI										
ManyDepth [Watson et al.]	2021	$2, (-1, 0)$	M	0.098	0.770	4.459	0.176	0.900	0.965	0.983
ProDepth [Woo et al.]	2024	$2, (-1, 0)$	M	0.086	0.629	4.139	0.166	0.918	0.969	0.984
DSI-MonoViT [Zhang et al.]	2025	1	M	0.096	0.711	4.321	0.172	0.905	0.968	0.984
ManyDepth2 [Zhou et al.]	2025	$2, (-1, 0)$	M+F	0.091	0.649	4.232	0.170	0.909	0.968	0.984
PuriLight [Chen et al.]	2025	1	M	0.106	0.747	4.536	0.182	0.890	0.964	0.984
DiMoDE [Zhang et al.]	2026	1	M	0.097	0.652	4.337	0.172	0.899	0.967	0.985
VFM-Dynamo [Du et al.]	2026	$2, (-1, 0)$	M+F	0.086	0.642	4.129	0.165	0.916	0.971	**0.986**
FlexDepth[Zhu et al.]	2026	1	M	0.093	0.605	4.114	0.167	0.910	0.969	0.985
FlexDepth++		$2, (-1, 0)$	M	**0.082**	**0.538**	**3.901**	**0.160**	**0.926**	**0.972**	0.985
Cityscapes										
ManyDepth [Watson et al.]	2021	$2, (-1, 0)$	M	0.114	1.193	6.223	0.170	0.875	0.967	0.989
DynamicDepth [Feng et al.]	2022	$2, (-1, 0)$	M+Se	0.103	1.000	5.867	0.157	0.895	0.974	0.991
MonoViT [Zhao et al.]	2022	1	M	0.114	1.238	6.589	0.174	0.860	0.965	0.990
ProDepth [Woo et al.]	2024	$2, (-1, 0)$	M	0.095	0.876	5.531	0.146	0.908	0.978	0.993
DSI-MonoViT [Zhang et al.]	2025	1	M	0.088	0.872	5.410	0.141	0.918	0.981	0.993
ManyDepth2 [Zhou et al.]	2025	$2, (-1, 0)$	M+F	0.097	**0.792**	5.827	0.154	0.903	0.975	0.993
PuriLight[Chen et al.]	2025	1	M	0.099	1.014	5.892	0.153	0.901	0.976	0.992
FlexDepth[Zhu et al.]	2026	1	M	0.086	0.877	5.268	0.137	0.926	0.982	0.993
FlexDepth++		$2, (-1, 0)$	M	**0.083**	0.794	**5.145**	**0.135**	**0.927**	**0.983**	**0.994**

Table 3: Ablation study of the proposed components on KITTI Geiger et al. (2013) and UAVScenes Wang et al. (2025c). **MCVC**: Multi-resolution Cost-Volume Construction; **SCVR**: Stability-guided Cost-Volume Refinement.

	MCVC	SCVR	KITTI			UAVScenes		
			Abs Rel↓	Sq Rel↓	$\delta < 1.25$ ↑	Abs Rel↓	Sq Rel↓	$\delta < 1.25$ ↑
(a)	×	×	0.084	0.554	0.923	0.890	0.925	0.936
(b)	✓	×	**0.082**	0.545	**0.926**	0.075	0.723	0.952
(c)	×	✓	**0.082**	0.552	0.925	0.072	0.761	0.951
(d)	✓	✓	**0.082**	**0.538**	**0.926**	**0.071**	**0.722**	**0.954**

of 0.075, a Sq Rel of 0.723, and a $\delta < 1.25$ of 0.952. These results indicate that constructing cost volumes from complementary feature resolutions provides richer temporal matching context and is particularly beneficial in aerial scenes with broader depth variation. Method **(c)** extends **(a)** by incorporating the Stability-guided Cost-Volume Refinement (SCVR) module. Compared with the baseline, SCVR also improves the overall performance, achieving an Abs Rel of 0.082 on KITTI and 0.072 on UAVScenes. In particular, the improvement on UAVScenes is clear, suggesting that the stability-guided refinement effectively suppresses unstable matching responses and enhances the reliability of temporal geometric cues under more challenging aerial motion patterns. Method **(d)** is the full FlexDepth++ model, which combines both MCVC and SCVR. By integrating the two modules, the model achieves the best overall performance on both benchmarks. On KITTI, it attains the lowest Sq Rel of 0.538 while maintaining the best Abs Rel of 0.082 and $\delta < 1.25$ of 0.926. On UAVScenes, it further improves the performance to an Abs Rel of 0.071, a Sq Rel of 0.722, and a $\delta < 1.25$ of 0.954. These results show that MCVC and SCVR are complementary: MCVC enhances the expressiveness of temporal matching, while SCVR further improves its stability, and their combination leads to the strongest performance.

Qualitative Results. Figures 3 and 4 show qualitative comparisons on UAVLight and KITTI. In low-altitude aerial scenes, FlexDepth++ generates depth maps that are more consistent with the ground truth, especially in regions with large depth variations and complex spatial layouts. Compared with Depth Anything V2, whose relative depth ordering is visually plausible, the aligned depth distribution is strongly distorted: e.g., in the first frame of Figure 3, the ground truth spans 58–84 m, whereas the median-aligned DA-V2 prediction expands to roughly 54–355 m. ManyDepth2 preserves the depth range better, but its predictions on thin structures such as trees deviate noticeably from the ground truth. On KITTI, FlexDepth++ also recovers sharper object boundaries and richer

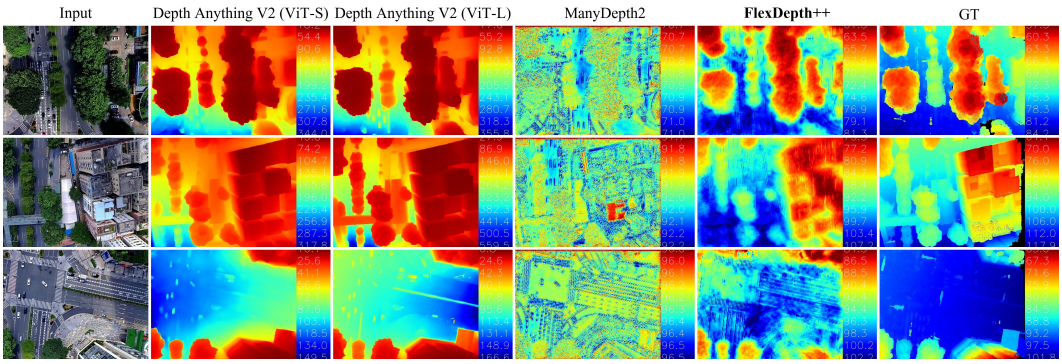


Figure 3: Qualitative comparison in UAV scenerios on UAVLight Du et al. (2026a). FlexDepth++ yields more accurate depth estimation compared with DepthAnything V2 Yang et al. (2024a) and ManyDepth2 Zhou et al. (2025). All predictions are median-scaled to the corresponding GT depth for visualization, with the color bars on the right indicate the metric depth scale of the GT.

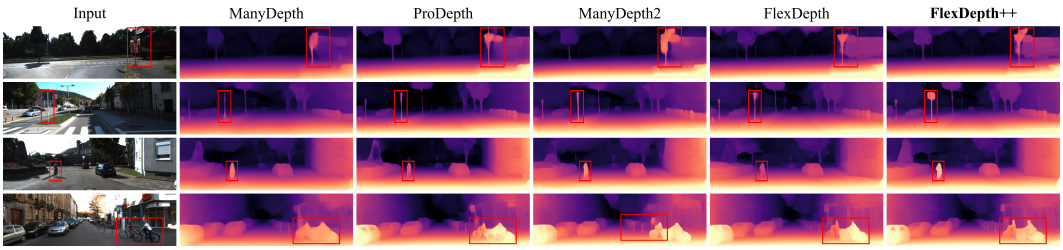


Figure 4: Qualitative comparison in autonomous driving scenerios on KITTI Geiger et al. (2013). The red boxes highlight representative static objects (the first two rows) and dynamic objects (the last two rows), where FlexDepth++ better preserves fine-grained boundaries and textures details compared with other mothods.

fine-grained details than other methods, showing that our method generalizes well to driving scenes in addition to aerial scenes. These results demonstrate that FlexDepth++ consistently improves geometric reconstruction quality across both low-altitude aerial and autonomous driving scenarios.

5 CONCLUSION

In this paper, we introduce *FlexDepth++*, a unified self-supervised multi-frame monocular depth estimation framework for autonomous driving and low-altitude aerial scenes. Unlike existing multi-frame methods that are primarily developed for structured driving environments, FlexDepth++ enhances temporal geometric matching through a hierarchical cost-volume construction-and-refinement paradigm. Multi-resolution Cost-Volume Construction aggregates complementary matching cues from different feature resolutions to enrich temporal context, while Stability-guided Cost-Volume Refinement leverages training-stage prediction stability to improve the robustness of cost-volume representations. By jointly addressing insufficient matching context and temporal matching instability, FlexDepth++ generalizes effectively across substantially different camera motions, depth distributions, and scene structures. Experiments on UAVScenes, UAVLight, KITTI, and Cityscapes show state-of-the-art performance across both aerial and driving benchmarks, demonstrating the effectiveness of the proposed unified framework.

Limitations. The richer temporal matching design in FlexDepth++ comes with a moderate increase in computational cost. Future work will explore more efficient cost-volume construction and refinement strategies to further improve the performance–efficiency trade-off. In addition, our current evaluation focuses on autonomous driving and low-altitude aerial scenes, and extending the framework to indoor and other heterogeneous environments remains an important direction.

REFERENCES

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*, volume 33, pp. 8001–8008, 2019.

Yujie Chen, Li Zhang, Xiaomeng Chu, and Tian Zhang. Purilight: A lightweight shuffle and purification framework for monocular depth estimation. In *ECAI 2025: 28th European Conference on Artificial Intelligence 25-30 October 2025, Bologna, Italy-Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS 2025) Proceedings*, pp. 3274–3281, 2025. doi: 10.3233/FAIA251195.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Kang Du, Xue Liao, Junpeng Xia, Chaozheng Guo, Yi Gu, Yirui Guan, Duotun Wang, Sheng Huang, and Zeyu Wang. Uavlight: A benchmark for illumination-robust 3d reconstruction in unmanned aerial vehicle (uav) scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5670–5679, June 2026a.

Qianqian Du, Hui Yin, Xingyu Miao, and Zhengyin Liang. Vfm-dynamo: Accurate non-rigid motion identification via vision foundation model for self-supervised monocular depth estimation. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 1052–1060. International Joint Conferences on Artificial Intelligence Organization, 2026b. doi: 10.24963/ijcai.2026/118.

Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In *Computer Vision – ECCV 2022*, pp. 228–244. Springer, 2022. doi: 10.1007/978-3-031-19824-3_14.

Xuanang Gao, Zhiwei Ning, Gengming Zhang, Jiaxi Cao, Runze Yang, Zhonglong Zheng, Jie Yang, Rong Xiao, and Wei Liu. Rosamdepth: Robust self-supervised depth estimation leveraging segment anything model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34138–34147, 2026.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3828–3838, 2019.

Leezy Han, Seunggyu Kim, Dongseok Shim, and Hyeonbeom Lee. Ptc-depth: Pose-refined monocular depth estimation with temporal consistency. *arXiv preprint arXiv:2604.01791*, 2026.

Shenghui Huang, Menghao Hu, Longkun Zou, Hongyu Chi, Zekai Li, Feng Gao, Fan Yang, Qingyao Wu, and Ke Chen. Uav-cb: A complex-background rgb-t dataset and local frequency bridge network for uav detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40468–40478, June 2026.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Zhenyu Li, Wenqing Cui, Shariq Farooq Bhat, and Peter Wonka. Patchrefiner v2: Fast and lightweight real-domain high-resolution metric depth estimation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=AAqeYdGdn2.

Logambal Madhuanand, Francesco Nex, and Michael Ying Yang. Self-supervised monocular depth estimation from oblique uav videos. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176: 1–14, 2021. doi: 10.1016/j.isprsjprs.2021.03.024.

Liang Qin, Min Wang, Xingyu Lu, Aowen Qiu, Wengang Zhou, and Houqiang Li. Uast: Unified active search and tracking for arbitrary targets with uavs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13464–13473, June 2026.

Zhongqiang Song, Guanying Chen, Yuqi Zhang, Yin Zou, Chuanyu Fu, Zhiyuan Yuan, Chuan Huang, Shuguang Cui, and Xiaochun Cao. Aerialmetric: Benchmarking and adapting uav monocular metric depth estimation in the real world. *arXiv preprint arXiv:2606.29716*, 2026.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5261–5271, June 2025a.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-1207.

Sijie Wang, Siqi Li, Yawei Zhang, Shangshu Yu, Shenghai Yuan, Rui She, Quanjiang Guo, JinXuan Zheng, Ong Kang Howe, Leonrich Chandra, et al. Uavscenes: A multi-modal dataset for uavs. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28946–28958. IEEE, 2025c.

Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1164–1174. IEEE, 2021.

Sungmin Woo, Wonjoon Lee, Woo Jin Kim, Dogyoon Lee, and Sangyoun Lee. Prodepth: Boosting self-supervised multi-frame monocular depth with probabilistic fusion. In *Computer Vision – ECCV 2024*, pp. 201–217. Springer, 2024. doi: 10.1007/978-3-031-72646-0_12.

Haifeng Wu, Shuhang Gu, Lixin Duan, and Wen Li. Geodepth: From point-to-depth to plane-to-depth modeling for self-supervised monocular depth estimation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11525–11535. IEEE, 2025a.

You Wu, Xucheng Wang, Xiangyang Yang, Mengyuan Liu, Dan Zeng, Hengzhou Ye, and Shuiwang Li. Learning occlusion-robust vision transformers for real-time uav tracking. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17103–17113. IEEE, 2025b.

Chaocan Xue, Qihua Liang, Bineng Zhong, Yanting Zu, Yuanliang Xue, Haiying Xia, and Shuxiang Song. Toward low-cost yet effective temporal learning for uav tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 42538–42548, June 2026.

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*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Jingtao Ye, Kexin Zhang, Xunchi Ma, Yuechan Li, Guangming Zhu, Peiyi Shen, Linhua Jiang, Xiangdong Zhang, and Liang Zhang. Breaking smooth-motion assumptions: A uav benchmark for multi-object tracking in complex and adverse conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13594–13603, June 2026.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

Mengtan Zhang, Zizhan Guo, Hongbo Zhao, Yi Feng, Zuyi Xiong, Yue Wang, Shaoyi Du, Hanli Wang, and Rui Fan. Discriminately treating motion components evolves joint depth and ego-motion learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3732069. Early Access.

Tian Zhang, Li Zhang, Xiaomeng Chu, Beibei Wang, Zhaowen Zhu, and Yu-An Liu. Depth self-inhibition: An advanced training framework for self-supervised monocular depth estimation in dynamic scenes. *TechRxiv*, 2025. doi: 10.36227/techrxiv.174001043.36290744/v1.

Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. MonoViT: Self-supervised monocular depth estimation with a vision transformer. In *2022 International Conference on 3D Vision (3DV)*, pp. 668–678. IEEE, 2022. doi: 10.1109/3DV57658.2022.00077.

Kaichen Zhou, Jia-Wang Bian, Jian-Qing Zheng, Jiaxing Zhong, Qian Xie, Niki Trigoni, and Andrew Markham. Manydepth2: Motion-aware self-supervised monocular depth estimation in dynamic scenes. *IEEE Robotics and Automation Letters*, 10(7):6704–6711, 2025.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 6612–6619. IEEE, 2017.

Zhaowen Zhu, Li Zhang, Yujie Chen, Tian Zhang, Yingjie Wang, and Mingxia Zhan. Towards robust driving perception: A flexible scale-driven family for self-supervised monocular depth estimation. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 110–129, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-36846-1. doi: 10.1007/978-3-032-36846-1_7.