Crafting Monocular Cues and Velocity Guidance for Self-Supervised Multi-Frame Depth Learning

Xiaofeng Wang\textsuperscript{1,2}, Zheng Zhu\textsuperscript{3}, Guan Huang\textsuperscript{3}, Xu Chi\textsuperscript{3}, Yun Ye\textsuperscript{3}, Ziwei Chen\textsuperscript{4}, Xingang Wang\textsuperscript{1*}

\textsuperscript{1} Institute of Automation, Chinese Academy of Sciences
\textsuperscript{2} School of Artificial Intelligence, University of Chinese Academy of Sciences
\textsuperscript{3} PhiGent Robotics
\textsuperscript{4} Southeast University

\{wangxiaofeng2020, xingang.wang\}@ia.ac.cn, zhengzhu@ieee.org
\{guan.huang, xu.chi, yun.ye\}@phigent.ai, richard.chen@seu.edu.cn

Abstract

Self-supervised monocular methods can efficiently learn depth information of weakly textured surfaces or reflective objects. However, the depth accuracy is limited due to the inherent ambiguity in monocular geometric modeling. In contrast, multi-frame depth estimation methods improve the depth accuracy thanks to the success of Multi-View Stereo (MVS), which directly makes use of geometric constraints. Unfortunately, MVS often suffers from texture-less regions, non-Lambertian surfaces, and moving objects, especially in real-world video sequences without known camera motion and depth supervision. Therefore, we propose MOVEDepth, which exploits the MONocular cues and VElocity guidance to improve multi-frame Depth learning. Unlike existing methods that enforce consistency between MVS depth and monocular depth, MOVEDepth boosts multi-frame depth learning by directly addressing the inherent problems of MVS. The key of our approach is to utilize monocular depth as a geometric priority to construct MVS cost volume, and adjust depth candidates of cost volume under the guidance of predicted camera velocity. We further fuse monocular depth and MVS depth by learning uncertainty in the cost volume, which results in a robust depth estimation against ambiguity in multi-view geometry. Extensive experiments show MOVEDepth achieves state-of-the-art performance: Compared with Monodepth2 and PackNet, our method relatively improves the depth accuracy by 20\% and 19.8\% on the KITTI benchmark. MOVEDepth also generalizes to the more challenging DDAD benchmark, relatively outperforming Many-Depth by 7.2\%. The code is available at https://github.com/JeffWang987/MOVEDepth.

Introduction

Depth estimation is a fundamental task in 3D computer vision, with versatile applications ranging from virtual/augmented reality (Luo et al. 2020) to autonomous driving (Geiger, Lenz, and Urtasun 2012). Although 3D sensors (e.g., LiDAR, structured light) can generate accurate depth information, it is more attractive to infer depth from a single RGB image in a self-supervised way (Garg et al. 2016; Godard, Aodha, and Brostow 2016; Godard et al. 2018), which eliminates the necessity of expensive 3D hardware and multi-sensor calibration. However, the accuracy of these monocular methods is not yet on par with 3D sensors due to their inherent ambiguity in geometric modeling.

To improve the monocular depth accuracy, recent multi-frame methods\textsuperscript{1} (Watson et al. 2021; Wimbauer et al. 2020; Wang, Pizer, and Frahm 2019; Zhang et al. 2019; Feng et al. 2022; Patil et al. 2020; Wang, Pang, and Lin 2022) leverage temporal and spatially associations in multi-frame video sequences which are available in real-world scenes (e.g., smart devices (Ha et al. 2016) or moving vehicles (Menze and Geiger 2015)). Among these approaches, cost-volume-based methods (Feng et al. 2022; Watson et al. 2021; Wimbauer et al. 2020) achieve state-of-the-art depth accuracy as they take advantage of the successful Multi-View Stereo (MVS). However, MVS is still challenged by unsatisfactory reconstructions in real-world scenes with non-Lambertian surfaces, textureless areas, and moving objects (Knapsch et al. 2017; Schöps et al. 2017). To tackle these problems, teacher-student training architectures (Watson et al. 2021; Feng et al. 2022; Shao et al. 2022) are proposed to enforce consistency between monocular depth and MVS depth. However, the consistency pushes MVS depth to mimic monocular depth, which underuses the multi-view geometry, thus the performance of these methods is limited.

To improve the multi-view depth accuracy, the learning-based MVS methods (Yao et al. 2018, 2019) densely sample depth candidates in a large range. However, the dense sampling strategy causes matching ambiguity in real-world video frames without known camera pose and depth supervision (see Fig. 3). To mitigate the problem, we explore an efficient approach to enhance geometric cues for improving self-supervised multi-frame depth learning. Our intuition is that the monocular depth serves as a geometric priority of the scene, and the multi-frame matching ambiguity can be significantly reduced by sampling depth candidates near the monocular priority. Apart from the matching ambiguity, the multi-view geometry is still challenged by insufficient Tri-

\textsuperscript{1}Corresponding author.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
angulation Prior (Schonberger et al. 2016), especially in static/slow video sequences where nearby frames share little stereo baseline. To address the problem, the predicted camera velocity is leveraged to adaptively adjust the depth range. Specifically, multi-frames with higher motion velocity have larger viewpoint change, which can benefit multi-view geometry, so the depth range is enlarged to infer more accurate depth. In contrast, static frames carry little information for depth inference, thus the depth range is shrunken to the more reliable monocular priority. Besides, we fuse monocular depth and MVS depth by learning uncertainty in the cost volume, resulting in a robust depth estimation against artifacts in multi-view geometry (e.g., moving objects, textureless areas).

Owing to the monocular depth priority and camera velocity guidance, MOVEDepth achieves state-of-the-art performance: Compared with competitive monocular baselines (Godard et al. 2018; Guizilini et al. 2019), our method relatively improves the depth accuracy by ~20% on the KITTI benchmark. MOVEDepth also generalizes to the more challenging DDAD benchmark, relatively outperforming Many-Depth (Watson et al. 2021) by 7.2%. Besides, qualitative analysis demonstrates that our method is more robust against challenging artifacts where multi-view geometry fails.

The main contributions are three-fold as follows:
- We propose a novel self-supervised multi-frame depth learning framework, named MOVEDepth. It leverages the monocular depth cues as a geometric priority, and the multi-frame matching ambiguity is mitigated by sampling depth candidates near the monocular priority.
- The velocity-guided depth sampling is proposed to address failure cases caused by slow/static camera motion. And an adaptive fusing layer is introduced to learn uncertainty in cost volume, which mitigates artifacts brought by textureless areas and moving objects.
- We conduct extensive experiments on KITTI and DDAD, and the results show our method achieves superior depth accuracy in the complex real-word scenes with fewer depth candidates.

Related Work

In this section, we review depth estimation approaches relevant to our method in the following two categories: monocular depth learning and multi-frame depth learning.

Monocular Depth Learning

Despite the inherently ill-posed problems of geometric reasoning, monocular depth learning has been studied extensively in the literature. Supervised methods exploit sparse annotations (Chen et al. 2016) or dense points clouds from 3D sensors (Eigen, Puhrsch, and Fergus 2014; Eigen and Fergus 2014; Fu et al. 2018). Self-supervised approaches mitigate the expensive 3D sensors and human annotations, instead training with photo-metric consistency losses using stereo images (Xie, Girshick, and Farhadi 2016; Garg et al. 2016). Recently, monocular video supervision (Zhou et al. 2017; Godard et al. 2018; Guizilini et al. 2019) is an attractive alternative to stereo-based methods but it brings a new set of challenges: The model needs to jointly estimate camera ego-motion and image depth to form reprojection losses, which further increases the difficulty of monocular geometric modeling. Although there still exists a huge performance gap between monocular methods and other geometric methods, recent advances in monocular depth learning provide a plausible geometric priority, which can benefit other geometric approaches to refine the depth.

Multi-Frame Depth Learning

Given multiple 2D images and camera parameters, MVS can reconstruct the dense geometry of the scene (Schonberger et al. 2016). Recent advances in learning-based MVS approaches (Yao et al. 2018; Gu et al. 2020; Wang et al. 2022b) further improve the reconstruction accuracy. However, these methods assume known camera poses and static scenes. MonoRec (Wimbauer et al. 2020) extends MVS in dynamic environments that are captured with a single moving camera, but it needs the pre-trained segmentation network (He et al. 2017) and sparse depth obtained by a visual odometry system (Yang et al. 2018). MaGNet (Bae, Budvytis, and Cipolla 2022) fuses single-view probability with multi-view geometry to produce robust estimations. However, it still requires ground truth depth supervision. The self-supervised method (Watson et al. 2021) removes the requirement for human annotations and depth supervision, and it proposes a teacher-student training architecture to encourage the network to ignore unreliable regions in MVS cost volume. (Feng et al. 2022) further improves the depth accuracy in regions of dynamic objects by disentangling object motions, and (Shao et al. 2022) utilizes Deformable Convolution Networks (DCNs (Dai et al. 2017)) to enhance the depth estimates in low-texture and homogeneous-texture regions. However, these methods enforce consistency between MVS depth and monocular depth, which underuses the geometric reasoning of MVS. In contrast, the proposed MOVEDepth directly addresses the inherent problems of MVS (e.g., static/slow camera motion, object motions, and textureless regions).

Method

A detailed description of MOVEDepth is given in this section and the main network architecture is illustrated in Fig. 1. Given a video sequence, (a) we firstly utilize PoseNet to estimate camera ego-motion and velocity between frame $T$ and frame $T-1$. (b) Then the monocular depth is predicted using the DepthNet. (c) Subsequently, we conduct homography warping between the encoded frame features using the predicted camera ego-motion and monocular depth priority. The resulting cost volume is decoded into a depth map and an uncertainty map, which serves as guidance for fusing the monocular depth and MVS depth. (d) Significantly, the depth sampling range of homography warping is adaptively adjusted under the guidance of predicted camera velocity, which mitigates problems brought by slow/static camera motion.

The following parts start with the preliminary of self-supervised monocular depth learning from a video sequence.
Figure 1: The main network architecture of MOVEDepth. (a) PoseNet is utilized to estimate camera ego-motion and velocity between frame $T$ and frame $T - 1$. (b) The monocular depth is predicted using the DepthNet, which serves as a geometric priority to construct MVS cost volume. (c) We conduct homography warping between the encoded feature maps using the predicted camera ego-motion and monocular depth priority. The resulting cost volume is decoded into a depth map and an uncertainty map. (d) The depth sampling range is adaptively adjusted under the guidance of predicted camera velocity.

We then introduce the important innovations to improve multi-frame depth learning by fusing monocular cues.

Self-Supervised Monocular Depth Learning

The self-supervised pipeline is conducted by jointly training a DepthNet $\theta_d$ (see Fig. 1(a)) and a PoseNet $\theta_p$ (see Fig. 1(b)) (Zhou et al. 2017), and they are trained only on video frames $\{I_t\}_{t=1}^{N}$. Specifically, we estimate the monocular depth $D_{\text{Mono}} \approx \theta_d(I_t)$ of current frame $I_t$, and predict relative camera pose $[R \mid T]_{t \rightarrow t+k} = \theta_p(I_t, I_{t+k})$ between frame $I_t$ and frame $I_{t+k}$ ($k \in -1, 1$). Then, we can synthetically $I_t$ from viewpoint $I_{t+k}$ by the following operation:

$$I_{t+k \rightarrow t}(D_{\text{Mono}}) = I_{t+k} \langle \text{proj}\big(D_{\text{Mono}} \cdot [R \mid T]_{t \rightarrow t+k} \cdot K\big) \rangle,$$

where $K$ is the camera intrinsics, $\text{proj}()$ is the projection function that returns the 2D pixel coordinates of the projected $D_{\text{Mono}}$, and $\langle \cdot \rangle$ is the pixel sampling operator. Following the optimization strategy (Godard et al. 2018), the training pipeline is optimized by a reprojection loss:

$$L_r(D_{\text{Mono}}) = \min_k pe\big(I_t, I_{t+k \rightarrow t}(D_{\text{Mono}})\big),$$

where the $\min$ operation selects the best matching pairs to avoid ambiguity brought by occlusions, and $pe(\cdot)$ is a weighted combination of $L_1$ loss and structure similarity (SSIM) loss. The reprojection loss is calculated over multi-scale depth outputs, and more implementation details can be found in (Godard et al. 2018).

MOVEDepth Design

MVS Depth from Monocular Depth Priority

Multi-view approaches warp source images into the reference camera frustum to form cost volume, and estimate depth to be the highest-activated value in cost volume (Yao et al. 2018). Although the hard-coded multi-view methods reduce geometry ambiguity and generate more accurate depth, they are still challenged by texture-less regions, non-Lambertian surfaces, and moving objects, especially in real-world video frames without known camera motion. The monocular methods, on the other hand, are more robust against weakly textured regions or moving objects but the overall depth accuracy is limited. Therefore, we exploit monocular cues to complement the limitation of MVS (see Fig. 1(c)), which is elaborated in the following.

Given a current frame $I_t \in \mathbb{R}^{H \times W \times 3}$ and its nearby frame $I_{t-1} \in \mathbb{R}^{H \times W \times 3}$ (the future frames is not used to enable online depth prediction), we firstly leverage an encoder $\theta_{\text{enc}}$ to extract 2D features of these frames, where the images are downsampled to lower resolution deep features $F_{t(i \in \{0, -1\})} \in \mathbb{R}^{H/4 \times W/4 \times C}$. Following previous learning-based MVS (Yao et al. 2018; Gu et al. 2020), the plane sweep stereo (Collins 1996) is utilized to establish multiple front-to-parallel planes in the current frame. Specifically, equipped with camera intrinsic $K$ and extrinsic $[R \mid T]$ estimated by PoseNet $\theta_p$, the previous frame features can be warped into the current camera frustum:

$$p_{t-1, j} = K \cdot \left( R \cdot (K^{-1} \cdot p_t \cdot d_j) + T \right),$$

where $p_t$ is the $j$-th pixel point in frame $t$, $d_j$ is the depth value of pixel $j$ in frame $t$, and $R$ and $T$ are the relative rotation and translation between frames $t-1$ and $t$, respectively. The depth range is automatically adjusted in the current frame. Finally, we use the pre-trained MVSNet (Yao et al. 2018) to further refine the predicted depth.
where \( d_j \) is the \( j \)-th hypothesized depth candidates of pixel \( p_j \) in the current frame feature \( F_i \), and \( d_{i-1,j} \) denotes the corresponding pixel in the previous frame feature \( F_{i-1} \). After the warping operation, the volume feature \( V_{i-1} \in \mathbb{R}^{H/4 \times W/4 \times C \times D} \) is constructed, where \( D \) is the number of depth candidates. Significantly, to reduce depth searching space, we specify the depth range \( \mathcal{R} \) using the monocural depth priority \( D_{\text{Mono}} \):

\[
\mathcal{R} = \{ d | d_{\text{min}} \leq d \leq d_{\text{max}} \},
\]

(4)

where \( (d_{\text{min}} + d_{\text{max}})/2 = D_{\text{Mono}}, \) and \( d_{\text{min}}, d_{\text{max}} \) is adaptively adjusted under the guidance of camera velocity, which is elaborated in the next subsection.

Given previous frame volume \( V_{i-1} \), we then use group correlation (Wang et al. 2021; Xu and Tao 2019) to construct cost volume, which measures the visual similarity between the current frame and the previous frame:

\[
s_i^g = \frac{1}{G} \langle v_i^g, f_i^g \rangle,
\]

(5)

where \( v_i^g \in \mathbb{R}^{H \times D} \) is the \( g \)-th group feature of \( v_i \) (\( v_i \in \mathbb{R}^{C \times D} \) is the \( i \)-th pixel feature of \( V_i \), and \( f_i^g \in \mathbb{R}^{C \times 1} \) is the \( g \)-th group feature of \( f_i \) (\( f_i \) is the \( i \)-th pixel feature of \( F_i \)), and \( \langle \cdot, \cdot \rangle \) is the inner product. Then \( \{ s_i^g \}_{g=0}^{G-1} \) are channel-wise stacked to generate \( s_i \in \mathbb{R}^{G \times D} \), which is \( i \)-th pixel feature of the final cost volume \( S_i \in \mathbb{R}^{H/4 \times W/4 \times G \times D} \).

The calculated cost volume is subsequently decoded by a light-weight \( \theta_{\text{dec}} \) to get depth probability \( P \in \mathbb{R}^{H/4 \times W/4 \times D} \), and the MVS depth is generated by \text{localmax} (Wang et al. 2022a):

\[
D_{\text{MVS}}(p) = \left( \frac{1}{\sum_{j=X(p)+r}^{X(p)+r} \frac{1}{d_j} \cdot p_j} \right)^{-1},
\]

(6)

where \( p \in \mathbb{R}^D \) is the pixel value of \( P, X(p) = \text{argmax}_p \), \( d_j \) is the index of the highest value for \( p \), and \( r \) is a radius parameter (typically set as 1). Finally, the \text{convex interpolation} (Teed and Deng 2020) is leveraged to upsample the MVS depth to the original resolution.

### Velocity-Guided Depth Sampling

In the previous subsection, the monocural depth is leveraged as a geometric center for depth sampling, but the depth range is left to be addressed. Typically, learning-based MVS (Yao et al. 2018, 2019) sample depth candidates in a fixed range, which is either calculated by COLMAP (Schonberger et al. 2016) or learned by networks (Watson et al. 2021). However, the depth range is utilized to describe the entire scene, and densely searching in such a wide range is computationally expensive and can not produce accurate depth (Gu et al. 2020). Recent methods reduce depth range by coarse-to-fine sampling (Gu et al. 2020; Wang et al. 2021) or confidence-based sampling (Bae, Budvytis, and Cipolla 2022; Cheng et al. 2020). However, we empirically find these sampling strategies are limited in self-supervised multi-frame depth learning (see Tab. 4), as they overlook the Triangulation Prior (Schonberger et al. 2016) of nearby frames.

To mitigate the problem, we propose velocity-guided depth sampling. The key innovation is to associate \text{Triangulation Prior} with camera motion velocity \( v \). Namely, the viewpoint changes noticeably when the camera moves at a high velocity, providing a sufficient \text{Triangulation Prior} for multi-view geometry. In contrast, slow/static video frames share a similar viewpoint, thus the \text{Triangulation Prior} is limited (theoretical analysis is in supplement). For video frames with insufficient \text{Triangulation Prior}, we expand depth range to infer accurate depth, and for frames with insufficient \text{Triangulation Prior}, the depth range is shrunk to the more reliable monocural priority. The depth sampling range is specified as follows:

\[
d_{\text{min}} = D_{\text{Mono}}(1 - \beta \| T \|_2)
\]

\[
d_{\text{max}} = D_{\text{Mono}}(1 + \beta \| T \|_2),
\]

(7)

where the camera motion velocity \( v = \alpha \| T \|_2 \) is the byproduct of PoseNet \( \theta_p \) (\( T \) is the camera translation estimated by \( \theta_p \), and \( \alpha \) is the camera frame rate), \( \beta \) is a hyper-parameter, and \( \| \cdot \|_2 \) is a scale function that transforms \( v \) to a real-world scale, which can be calculated by median-scaling (Godard et al. 2018) or camera-height-scaling (Yin et al. 2017). To ensure training stability, \( \beta \| T \|_2 \) is clamped to range \((0, 1)\).

Notably, the depth sampling strategy resembles \text{Gaussian Sampling} with mean of \( D_{\text{Mono}} \) and variance of \( \beta \| T \|_2 \), and the MVS depth range shrinks to the more reliable monocural depth when the camera is static. Differently, the depth candidates are not sampled by their probability, but by a deterministic inverse sampling strategy:

\[
d_j = \left( \frac{1}{d_{\text{min}}} - \frac{1}{d_{\text{max}}} \right) \frac{j}{D-1} + \frac{1}{d_{\text{max}}} \right)^{-1},
\]

(8)

where \( j = 0 \ldots D - 1 \). Compared with linear sampling (Yao et al. 2018) or probabilistic sampling (Bae, Budvytis, and Cipolla 2022), inverse depth sampling results in uniformly distributed depth candidates at the pixel level, which is beneficial for large-scale multi-frame matching (Xu and Tao 2019).

### Uncertainty-Based Depth Fusing

The calculated \( D_{\text{MVS}} \) is still challenged by texture-less regions, non-Lambertian
surfaces, and moving objects, which are the inherent problems of the multi-view geometry. To alleviate the problem, the uncertainty-based fusing method is introduced to replace unsatisfactory $D_{\text{MVS}}$ with the more reliable $D_{\text{Mono}}$. As shown in Fig. 2, we leverage an Uncertainty Decoder $\theta_u$ to learn uncertainty map $U$ from the entropy of depth probability $p$:

$$U(p) = \theta_u \left( \sum_{j=0}^{D-1} -p_j \log p_j \right), \quad (9)$$

where $\theta_u$ is comprised of 2D Convolutional Neural Network (CNN) blocks and a Sigmoid function. The reason for adopting the entropy is that the randomness of the depth probability distribution is positively related to the MVS depth uncertainty (Zhang et al. 2020). Subsequently, the uncertainty map is leveraged to calculate the fused depth $D_{\text{Fuse}}$:

$$D_{\text{Fuse}} = U \odot D_{\text{Mono}} + (1 - U) \odot D_{\text{MVS}}, \quad (10)$$

where $\odot$ denotes element-wise product.

**Loss Function** MOVEdth is end-to-end trained in a self-supervised manner, and the loss consists of three parts:

$$L_{\text{MOVEDepth}} = \lambda_1 L(D_{\text{Mono}}) + \lambda_2 L(D_{\text{MVS}}) + \lambda_3 L(D_{\text{Fuse}}), \quad (11)$$

where $\lambda_1, \lambda_2, \lambda_3$ are the loss weights, and $L(\cdot)$ is a weighted combination of reprojection loss $L_r$ (Eq. (2)) and depth smooth loss $L_u$ (Godard, Aodha, and Brostow 2016):

$$L(D) = L_r(D) + \gamma L_u(D), \quad (12)$$

where $\gamma$ denotes the loss weight.

**Experiment**

**Datasets**

MOVEdth is evaluated on KITTI (Geiger, Lenz, and Urtasun 2012) and DDAD (Guizilini et al. 2019) to verify the effectiveness. KITTI is an outdoor dataset in the driving scenario, which is the standard benchmark for depth evaluation. Following from the Eigen split (Eigen and Fergus 2014), with data preprocessing from (Zhou et al. 2017), the data is divided into 39810/4424/697 training, validation and test images. For DDAD, it is a diverse dataset of the high-curvature urban, and residential scenes curated by self-driving cars, which is a novel benchmark for depth evaluation. Notably, DDAD evaluates with longer distance and denser LiDAR ground-truth, which is particularly challenging for multi-frame methods. Following (Guizilini et al. 2019), only front-view images are used, resulting in 12560/3950 training and validation images.

**Implementation Details**

Following previous work (Godard et al. 2018; Watson et al. 2021), we use color-jitter and flip as training-time augmentations, and MOVEdth is trained with an input resolution of 640×192 (KITTI) and 640×384 (DDAD). We only use two frames $\{I_{t-1}, I_t\}$ for cost volume construction, and use $\{I_{t-1}, I_t, I_{t+1}\}$ for reprojection loss. We train MOVEdth for 20 epochs and optimize it with Adam (Kingma and Ba 2015). MOVEdth is trained on 4 NVIDIA RTX 3090 GPUs with batch size 6 on each GPU. The learning rate is initially set as 0.0002, which decays by a factor of 10 for the final 5 epochs. The feature extractor in $\theta_d$ and $\theta_p$ follows the same architecture in (Godard et al. 2018; Guizilini et al. 2019), and $\theta_{\text{enc}}, \theta_{\text{dec}}$ comprise light-weight 2D CNNs and 3D CNNs (more network details are in supplement). Following (Godard et al. 2018; Watson et al. 2021), the loss weight $\gamma$ is set as 0.001, and $\lambda_{\{i\in\{1,2,3\}\}} = 1$. For MVS cost volume construction, the number of depth candidates is 16, group correlation $G = 16$, and $\beta = 0.15$.

**KITTI Results**

We compare MOVEdth with monocular methods and multi-frame methods, and the quantitative results are shown in Tab. 1, where MOVEdth achieves a state-of-the-art performance among all the competitors. We observe that MOVEdth relatively improves the best-performing monocular method (Guizilini et al. 2020) by 12.7\% (on Abs. Rel.), which demonstrates it is significant to investigate multi-frame geometry to improve depth accuracy. Besides, MOVEdth relatively outperforms multi-frame methods (ManyDepth (Watson et al. 2021) and DynamicDepth (Feng et al. 2022)) by 9.2\% and 7.3\% (on Abs. Rel.), which verifies that our method is effective for multi-frame depth learning. Furthermore, we implement MOVEdth with depth priority from different monocular methods, and our method relatively improves Monodepth2 (Godard et al. 2018) and PackNet (Guizilini et al. 2019) by 20\% and 19.8\% (on Abs. Rel.), which verifies that MOVEdth benefits from advances in existing monocular methods. Notably, the resolution of depth priority is $\frac{1}{4}$ of the inputs, and the depth priority significantly reduces the depth candidates of MVS, thus MOVEdth (with Monodepth2) still runs at an acceptable speed (32 FPS) at single RTX 3090.

Qualitative results are presented in Fig. 3, where the MVS baseline (without monocular priority) fails in video-based depth learning due to matching ambiguity in a large depth range. In contrast, monocular method (Godard et al. 2018) and MOVEdth can generate plausible depth prediction. We further visualize the absolute relative depth error, which shows that MOVEdth can significantly reduce the depth error brought by monocular methods.

**DDAD Results**

DDAD is a novel benchmark for depth estimation, which evaluates depth with longer distance and denser LiDAR points. We conduct experiments on DDAD to verify our method can generalize to more challenging scenarios. As shown in Tab. 2, MOVEdth still outperforms previous competitors. Notably, our method relatively improves the state-of-the-art multi-frame method (Watson et al. 2021) by 7.6\% (on Abs. Rel.).

**Ablation Study**

In this subsection, we conduct an ablation study on the KITTI dataset to analyze the effectiveness of each component. Firstly, we analyze the impact of the monocular
priority for MVS, demonstrating that the monocular priority significantly reduces multi-frame matching ambiguity with fewer depth candidates. Then we verify that, compared with cascade sampling (Gu et al. 2020) and confidence sampling (Bae, Budvytis, and Cipolla 2022), multi-frame depth learning benefits more from the proposed velocity-guided depth sampling. Lastly, we qualitatively show that the uncertainty-based fusing makes MOVEDepth more robust against challenging artifacts where multi-view geometry fails (e.g., moving objects). In the following ablation study, unless specified, we utilize Monodepth2 (Godard et al. 2018) as the monocular priority, and only two frames (0, -1) are leveraged to construct MVS cost volume.

### Monocular Priority for Multi-Frame Depth Learning

As shown in Fig. 3, due to matching ambiguity in a large depth range, the multi-view geometry (without monocular priority) fails to learn multi-frame depth in a self-supervised manner. Besides, quantitative results (Tab. 3) also show that the MOVEDepth (without monocular priority) is inferior to the monocular priority (Godard et al. 2018), yet with $3 \times 4 \times$ GPU memory consumption.

When integrated with the monocular depth priority, the geometric uncertainty is significantly reduced, and MOVEDepth can produce superior depth predictions with less depth bin candidates. We empirically select 16 depth bins for MOVEDepth, as it strikes a balance between accuracy and memory consumption, and MOVEDepth with 16 depth bins relatively outperform the monocular baseline by 20% (on Abs. Rel.).

### Velocity-Guided Depth Sampling

As shown in Tab. 4, MOVEDepth with fixed depth sampling range (row 1,2) shows restricted performance, and increasing input frames (row 3,4) can not improve depth accuracy due to the inherent geometric ambiguity (without pose and depth supervision). Recent methods adjust depth range by cascade sampling (Gu et al. 2020) (half the depth range in the next stage) or confidence-based sampling (Bae, Budvytis, and Cipolla 2022) (adjust depth range according to monocular depth confidence). However, these sampling strategies (row 5,6) bring marginal depth accuracy improvement, as they overlook the Triangulation Prior nearby frames. In contrast,

---

**Table 1:** Comparison of MOVEDepth to existing self-supervised methods on the KITTI (Geiger, Lenz, and Urtasun 2012) Eigen split. At top we compare MOVEDepth with monocular methods using one frame at test time. At middle we compare MOVEDepth with multi-frame methods with multiple frames inputs at test time. At bottom we show our method with different monocular depth priority.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test frames</th>
<th>W × H</th>
<th>Abs. Rel.</th>
<th>Sq. Rel.</th>
<th>RMSE</th>
<th>RMSEP&lt;sub&gt;log&lt;/sub&gt;</th>
<th>δ &lt; 1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPC++ (Luo et al. 2019)</td>
<td>1</td>
<td>832 × 256</td>
<td>0.141</td>
<td>1.029</td>
<td>5.350</td>
<td>0.216</td>
<td>0.816</td>
</tr>
<tr>
<td>Struct2depth (M) (Casser et al. 2019)</td>
<td>1</td>
<td>416 × 128</td>
<td>0.141</td>
<td>1.026</td>
<td>5.291</td>
<td>0.215</td>
<td>0.816</td>
</tr>
<tr>
<td>Videos in the wild (Gordon et al. 2019)</td>
<td>1</td>
<td>416 × 128</td>
<td>0.128</td>
<td>0.959</td>
<td>5.230</td>
<td>0.212</td>
<td>0.845</td>
</tr>
<tr>
<td>Monodepth2 (Godard et al. 2018)</td>
<td>1</td>
<td>640 × 192</td>
<td>0.115</td>
<td>0.903</td>
<td>4.863</td>
<td>0.193</td>
<td>0.877</td>
</tr>
<tr>
<td>PackNet-SFM (Guizilini et al. 2019)</td>
<td>1</td>
<td>640 × 192</td>
<td>0.111</td>
<td>0.785</td>
<td>4.601</td>
<td>0.189</td>
<td>0.878</td>
</tr>
<tr>
<td>R-MSSF (Zhou et al. 2021)</td>
<td>1</td>
<td>1024 × 320</td>
<td>0.108</td>
<td>0.748</td>
<td>4.481</td>
<td>0.179</td>
<td>0.893</td>
</tr>
<tr>
<td>RM-Depth (Hui 2022)</td>
<td>1</td>
<td>640 × 192</td>
<td>0.108</td>
<td>0.710</td>
<td>4.513</td>
<td>0.183</td>
<td>0.884</td>
</tr>
</tbody>
</table>

| GLNet (Chen et al. 2019)      | 3 (-1.0, +1) | 416 × 128 | 0.099 | 0.796 | 4.743 | 0.186 | 0.884 |
| CoMoDA (Kuznetsova et al. 2021) | N           | 640 × 192 | 0.103 | 0.862 | 4.594 | 0.183 | 0.899 |
| Patil (Patil et al. 2020)     | N           | 640 × 192 | 0.111 | 0.821 | 4.650 | 0.187 | 0.883 |
| TC-Depth (Rukkamp et al. 2021) | 3 (-1.0, +1) | 640 × 192 | 0.103 | 0.746 | 4.483 | 0.185 | 0.894 |
| ManyDepth (Watson et al. 2021) | 2 (-1.0)    | 640 × 192 | 0.098 | 0.770 | 4.459 | 0.176 | 0.900 |
| DynamicDepth (Feng et al. 2022) | 2 (-1.0)    | 640 × 192 | 0.096 | 0.720 | 4.458 | 0.175 | 0.897 |

| MOVEDepth with Monodepth2 | 2 (-1.0) | 640 × 192 | 0.092 | 0.686 | 4.332 | 0.173 | 0.904 |
| MOVEDepth (with PackNet)  | 2 (-1.0) | 640 × 192 | 0.089 | 0.663 | 4.216 | 0.169 | 0.904 |

**Table 2:** Depth evaluation results on DDAD validation set (Guizilini et al. 2019). The above methods are evaluated for depth range of 200m without cropping.

<table>
<thead>
<tr>
<th>Method</th>
<th>Abs. Rel.</th>
<th>Sq. Rel.</th>
<th>δ &lt; 1.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monodepth2</td>
<td>0.213</td>
<td>4.975</td>
<td>0.761</td>
</tr>
<tr>
<td>Packnet</td>
<td>0.162</td>
<td>3.917</td>
<td>0.823</td>
</tr>
<tr>
<td>GUDA</td>
<td>0.147</td>
<td>2.922</td>
<td>0.809</td>
</tr>
<tr>
<td>ManyDepth</td>
<td>0.145</td>
<td>3.246</td>
<td>0.821</td>
</tr>
<tr>
<td>Ours (w/ Monodepth2)</td>
<td>0.136</td>
<td>3.027</td>
<td>0.835</td>
</tr>
<tr>
<td>Ours (w/ PackNet)</td>
<td>0.134</td>
<td>2.903</td>
<td>0.837</td>
</tr>
</tbody>
</table>

**Table 3:** Ablation analysis on monocular priority for multi-frame depth learning, including depth metrics and inference GPU memory consumption (GPU consumption is measured by torch.cuda.max_memory_allocated).
Figure 3: Qualitative results on KITTI. We visualize depth prediction from (a) MVS baseline (MOVEDepth without monocular priority), (b) monocular method (Godard et al. 2018), and (c) MOVEDepth. (d)(e) We also visualize the absolute relative error (Abs. Rel.) compared to the ground truth, where the error ranges from blue (Abs. Rel. = 0.0) to red (Abs. Rel. = 0.2).

Table 4: Ablation analysis on velocity-guided depth sampling. MOVEDepth† denotes fixed depth range \([\frac{1}{3}D_{Mono}, \frac{2}{3}D_{Mono}]\), and MOVEDepth‡ denotes fixed depth range \([\frac{2}{3}D_{Mono}, \frac{5}{3}D_{Mono}]\).

Table 5: Ablation analysis on uncertainty-based fusing.

Figure 4: Visualization of the uncertainty map, where the left figures are RGB images, and the right figures are the learned uncertainty maps (white: certain, black: uncertain).

Conclusion

In this paper, we propose MOVEDepth, which crafts monocular cues and velocity guidance for improving multi-frame depth learning in a self-supervised manner. The proposed method leverages monocular depth priority to reduce multi-frame matching geometry ambiguity. Specifically, MOVEDepth constructs cost volume with depth range centered at the monocular priority, and the depth range is adaptively adjusted by the predicted camera velocity. Therefore, video frames with insufficient Triangulation Prior shrink the depth range to monocular priority, and those with sufficient Triangulation Prior expand the depth range to infer more accurate depth. Moreover, we learn depth uncertainty from the entropy of MVS depth probability, which results in a more robust depth against challenging artifacts where multi-view geometry fails. Extensive experiments on KITTI and DDAD show that MOVEDepth achieves state-of-the-art depth accuracy with significantly reduced depth range. We hope that MOVEDepth can inspire more self-supervised multi-frame depth learning methods in the future.
References

Bae, G.; Budvytis, I.; and Cipolla, R. 2022. Multi-View Depth Estimation by Fusing Single-View Depth Probability with Multi-View Geometry. In CVPR.


Godard, C.; Aodha, O. M.; and Brostow, G. J. 2016. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In CVPR.

Godard, C.; Aodha, O. M.; Firman, M.; and Brostow, G. J. 2018. Digging Into Self-Supervised Monocular Depth Estimation. In ICCV.


Guizilini, V.; Ambrus, R.; Pillai, S.; Raventos, A.; and Gaidon, A. 2019. 3D Packing for Self-Supervised Monocular Depth Estimation. In CVPR.


Ha, H.; Im, S.; Park, J.; Jeon, H.-G.; and Kweon, I. S. 2016. High-Quality Depth from Uncalibrated Small Motion Clip. In CVPR.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2017. Mask R-CNN. In ICCV.


Knapsitsch, A.; Park, J.; Zhou, Q.; and Koltun, V. 2017. Tanks and temples: benchmarking large-scale scene reconstruction. ToG.

Kuznietsov; Yevhen; Proesmans; Marc; and Van Gool, L. 2021. CoMoDA: Continuous Monocular Depth Adaptation Using Past Experiences. In WACV.

Luo, C.; Yang, Z.; Wang, P.; Wang, Y.; Xu, W.; Nevatia, R.; and Yuille, A. 2019. Every pixel counts++: Joint learning of geometry and motion with 3D holistic understanding. PAMI.


Teed, Z.; and Deng, J. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In ECCV.


