IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Manydepth2: Motion-Aware Self-Supervised Monocular Depth Estimation in Dynamic Scenes

Kaichen Zhou, Jia-Wang Bian, Jian-Qing Zheng, Jiaxing Zhong, Qian Xie, Niki Trigoni, and Andrew Markham

***Abstract*—Despite advancements in self-supervised monocular depth estimation, challenges persist in dynamic scenarios due to the dependence on assumptions about a static world. In this paper, we present Manydepth2, to achieve precise depth estimation for both dynamic objects and static backgrounds, all while maintaining computational efficiency. To address the challenges introduced by dynamic content, we incorporate optical flow into monocular depth estimation, allowing our model to distinguish between dynamic and static regions in multi-frame inputs. We then construct a motion-aware cost volume across multiple frames by incorporating dynamic region information, which is used for accurate depth estimation. Furthermore, to improve the accuracy and robustness of the network architecture, we propose an attention-based depth network that effectively integrates information from feature maps at different resolutions by incorporating both channel and non-local attention mechanisms. Compared to methods with similar computational costs, Manydepth2 achieves a significant reduction of approximately five percent in root-mean-square error for self-supervised monocular depth estimation on the KITTI-2015 dataset.**

***Index Terms*—SLAM, visual-inertial SLAM, visual learning.**

I. Introduction

THE role of vision-based depth estimation (VDE) has become increasingly important in computer vision due to its ability to understand the 3D geometry of a scene based on 2D observation, which serves as the foundation for various high-level 3D tasks, such as scene reconstruction [1], object detection [2] and navigation [3]. Moreover, VDE has enabled state-of-the-art applications ranging from autonomous driving [3] to augmented reality [4], [5].

Recently, self-supervised depth estimation has emerged as a viable approach for training depth estimation methods, aiming to alleviate the dependency on extensive training data and reduce high computational demands. These methods learn depth maps from either monocular images [6] or stereo image pairs [7]. Despite significant progress in self-supervised monocular depth estimation, a substantial performance gap remains between self-supervised monocular and stereo methods, particularly in accurately estimating depth in dynamic regions. The disparity in performance can be mainly ascribed to the capability of stereo methods to utilize multiple views for constructing a feature volume, thereby incorporating a greater amount of 3D camera frustum information. While multi-frame monocular video depth estimation (VDE) methods such as [8] can construct a feature volume using adjacent frames—treating the current frame as the target and another as the reference—the presence of dynamic objects in these frames may disrupt the accuracy and consistency of the constructed feature volume.

Taking the aforementioned challenges into account, we propose ManyDepth2, a self-supervised multi-frame monocular depth estimation system. ManyDepth2 leverages a foundation optical flow model to distinguish between dynamic and static regions in the scene. By incorporating dynamic motion information, ManyDepth2 explicitly handles dynamic regions in the reference frame and constructs a motion-aware representation for matching pixels across the reference and target frames. This is achieved through an attention-based mechanism to build a motion-aware cost volume, improving depth estimation in dynamic environments. The attention mechanism is chosen due to its demonstrated outstanding performance in representation learning and effective fusion of diverse information. Experimental results indicate that our method could effectively handle dynamic objects, as demonstrated in Fig. 1.

- We utilize estimated optical flow alongside prior depth information to generate a pseudo-static reference frame. This reference frame effectively neutralizes the influence of dynamic elements within the original frame.
- By incorporating the pseudo-static reference frame, the target frame, and the initial reference frame, we construct a novel motion-aware volume that captures the dynamics of moving objects.
- Building on the High-Resolution Network (HRNet), we propose a novel depth estimation architecture that combines non-local and channel attention, enabling the integration of multi-scale features for accurate pixel-wise dense predictions.
- Our proposed model outperforms existing single and multi-frame methods on the KITTI, Cityscapes, and Odometry datasets. Additionally, our model can be trained efficiently

Received 27 January 2025; accepted 13 April 2025. Date of publication 8 May 2025; date of current version 26 May 2025. This article was recommended for publication by Associate Editor X. Chen and Editor J. Civera upon evaluation of the reviewers' comments. This work was supported in part by Amazon Web Services in the Oxford-Singapore Human-Machine Collaboration Programme and in part by the ACE-OPS project under Project EP/S030832/1. *(Corresponding author: Andrew Markham.)*

Kaichen Zhou, Jia-Wang Bian, Jian-Qing Zheng, Jiaxing Zhong, Niki Trigoni, and Andrew Markham are with the University of Oxford, OX1 2JD Oxford, U.K. (e-mail: zhouk777@mit.edu; andrew.markham@cs.ox.ac.uk).

Qian Xie is with the University of Leeds, LS2 9JT Leeds, U.K.

The code could found https://github.com/kaichen-z/Manydepth2.

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2025.3568337, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3568337

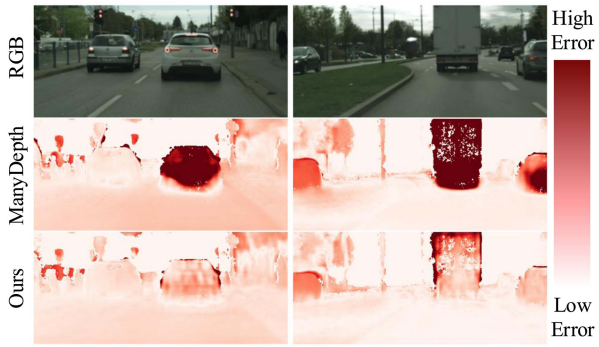

Fig. 1. Qualitative comparison on Cityscapes. The first row presents the RGB images. The second row shows the error maps of the depth estimates produced by ManyDepth. The third row displays the error maps of the depth estimates generated by Manydepth2.

using only a single NVIDIA RTX 3090 graphics card within a reasonable timeframe.

II. Related Works

A. Monocular Depth Estimation

Broadly, monocular VDEs fall into two categories: one relies on the present frame for depth estimation, while the other employs multiple frames to achieve depth prediction. **Single-frame Monocular Depth Estimation**: A prime example from this category is Monodepth, as highlighted in the study by [9]. This model distinguishes itself with a bifurcated network architecture: one part is dedicated explicitly to pose estimation, while another part focuses on depth estimation tasks. A pivotal aspect of the Monodepth architecture lies in the collaborative synergy between these two networks. They operate by leveraging the warping connection intrinsic to depth and image transformations. **Multi-frame Monocular Depth Estimation**: The premise of the second category rests on the notion that integrating temporal information during inference—by employing multiple neighboring frames as inputs—can enhance the accuracy of the final depth estimation. Initially, this is accomplished by employing test-time refinement techniques [10], along with recurrent neural networks (RNNs) as evidenced in studies like [11]. The test-time refinement method adopts a monocular strategy to leverage temporal data during testing, whereas the recurrent neural network integrates with a monocular depth estimation network to analyze continuous frame sequences. Nonetheless, models utilizing recurrent neural networks often entail high computational costs and lack a distinct geometric inference approach. Recently, Manydepth [8] and MonoRec [12] have made notable advancements in performance and real-time efficiency by incorporating cost volumes from stereo-matching tasks for geometric-based reasoning [13]. These models rely on a photometric loss function, where temporally neighboring frames are mapped onto the current image plane using predetermined depth bins. Within the cost volume framework, the inferred depth with the minimal value corresponds closely to the actual depth. Nevertheless, these approaches are grounded in static assumptions regarding scenarios and struggle with dynamic foreground elements. To tackle this limitation, we introduce Manydepth2, a technique adept at managing dynamic foreground by integrating temporal data into the cost volume and implementing a motion-aware photometric loss function.

B. Self-Supervised VDE for Dynamic Objects

Owing to the unique attributes of dynamic objects, there has been a focused effort among researchers to handle these objects distinctively throughout both the training and inference stages within self-supervised vision-based depth estimation (VDE) techniques. In [14], a deliberate decision was made to exclude dynamic objects from the analysis. This exclusion was aimed at preventing any potential interference with the optimization process, thereby ensuring the precision and accuracy of the depth estimation results. On the other hand, in [15], a more nuanced approach was adopted. Dynamic objects were first identified and segmented from the scene. Subsequent to this segmentation, these objects were treated distinctively within the framework of the photometric loss function. Such a strategy facilitates the integration of dynamic objects into the depth learning mechanism, ensuring that they contribute meaningfully without merely being sidelined or excluded. Furthermore, recent research efforts, as seen in [16], have ventured into predicting motion information at the object level. This predictive capability is then harnessed to establish more refined constraints, thereby enhancing the efficacy of self-supervised depth learning methodologies. Despite these advancements, it's noteworthy that many of these approaches either grapple with intensive computational requirements or demonstrate only marginal enhancements in performance metrics. Recognizing these challenges, our novel approach, termed Manydepth2, seeks to overcome these constraints. We introduce a motion-aware cost volume framework that seamlessly integrates forecasted optical flow insights, all achieved through a sophisticated attention mechanism.

III. Methodology

A. Overview

Firstly, as a foundational step, we utilize the potential of both optical flow information and pre-training depth data. By combining these insights, we construct a pseudo-static reference frame. This innovative approach ensures that the depth estimation process remains robust and relevant, even when confronted with dynamic elements within the scene. Our methodology further integrates the pseudo-static reference frame alongside the original target frame and vanilla reference frame to create a carefully designed motion-aware cost volume. This volume could improve the depth estimation by considering the detailed movement and space relationships in the scene. Finally, We are introducing a depth net structure that utilizes attention mechanisms for processing cost volume. This structure is meticulously crafted to effectively combine feature maps that come from different resolutions, all originating from a high-resolution network.

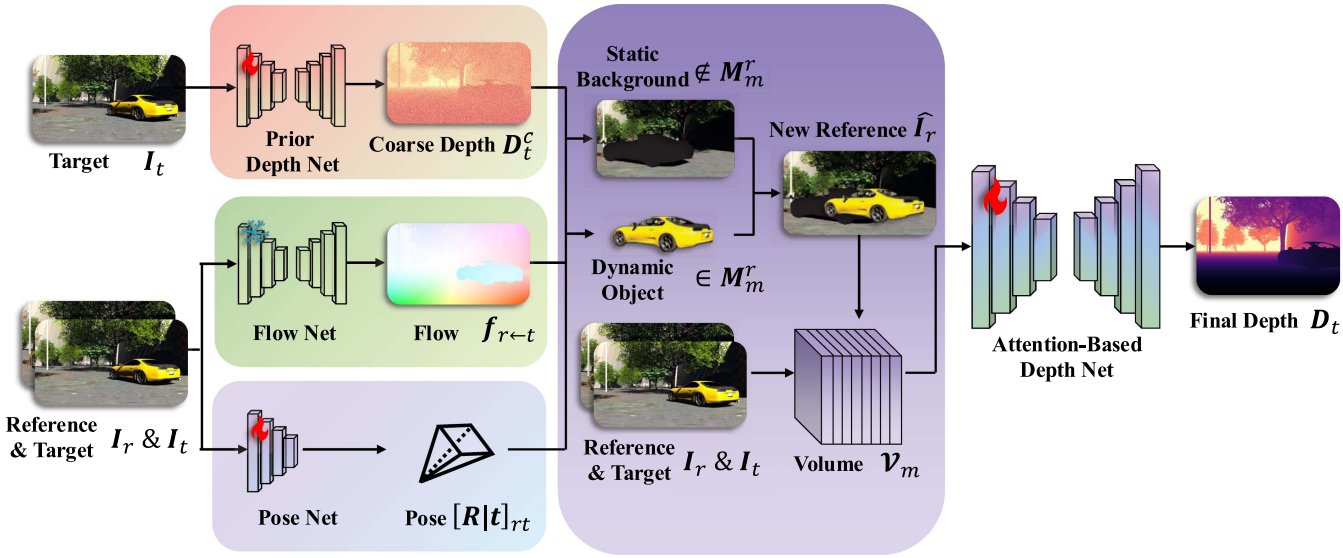

Fig. 2. Illustration for the structure of Manydepth2. During Stage 1, target and reference frame $\boldsymbol{I}_t$ and $\boldsymbol{I}_r$ are processed Flow Net, Pose Net, and Prior Depth Net to generate optical flow $\boldsymbol{f}_{r\leftarrow t}$, transformation matrix $[\boldsymbol{R}|\boldsymbol{t}]_{rt}$, and coarse depth $\boldsymbol{D}_t^c$. During Stage 2, the outputs $\boldsymbol{f}_{r\leftarrow t}$&$[\boldsymbol{R}|\boldsymbol{t}]_{rt}$&$\boldsymbol{D}_t^c$ are used to separate dynamic object and static background. A new reference frame $\hat{\boldsymbol{I}}_r$ is constructed based on (4). Then the motion-aware cost volume $\boldsymbol{\mathcal{V}}_m$ is constructed with the help of $\hat{\boldsymbol{I}}_r$&$\boldsymbol{I}_r$&$\boldsymbol{I}_t$&$[\boldsymbol{R}|\boldsymbol{t}]_{rt}$ and (5). Finally, the motion-aware cost volume $\boldsymbol{\mathcal{V}}_m$ and the target frame $\boldsymbol{I}_t$ are used by Attention-Based Depth Net to produce the refined depth $\boldsymbol{D}_t$.

B. Framework

Our objective is to estimate the depth map $\boldsymbol{D}_t$ and the rigid transformation $[\boldsymbol{R}|\boldsymbol{t}]_{rt}$ between the target frame $\boldsymbol{I}_t$ and a reference frame $\boldsymbol{I}_r$. To create the constraint for self-supervision, we can first reconstruct target frame $\boldsymbol{I}_t'$ in the following manner:

$$\boldsymbol{I}_t' = \phi(\boldsymbol{I}_r, \boldsymbol{D}_t, \boldsymbol{K}_t, \boldsymbol{K}_r, [\boldsymbol{R}|\boldsymbol{t}]_{rt}). \tag{1}$$

Here, $\boldsymbol{K}_t, \boldsymbol{K}_r$ represents the intrinsic matrices of the target and reference frames, and $\phi(.)$ denotes the projection process based on the image warping. The ultimate photometric loss function, denoted as $\mathcal{L}_p$, is computed as the discrepancy between the transformed image $\boldsymbol{I}_t'$ and the reference image $\boldsymbol{I}_t$. This loss function serves as the optimization objective for the proposed neural network.

The structure of Manydepth2 is demonstrated in Fig. 2. At the outset, Manydepth2 employs a trained optical flow network denoted by $\boldsymbol{\theta}_f$ and a pose network denoted by $\boldsymbol{\theta}_p$ to predict the optical flow $\boldsymbol{f}_{r\leftarrow t}$ and the transformation matrix $[\boldsymbol{R}|\boldsymbol{t}]_{rt}$ from the target frame to reference frame. Concurrently, a prior depth network denoted by $\boldsymbol{\theta}_{pd}$ is utilized to estimate the coarse depth map $\boldsymbol{D}_t^c$ for the target frame. Subsequently, the target frame denoted as $\boldsymbol{I}_t$, the reference frame represented by $\boldsymbol{I}_r$, the coarse depth map denoted as $\boldsymbol{D}_t^c$, the transformation matrix $[\boldsymbol{R}|\boldsymbol{t}]_{rt}$, and the optical flow $\boldsymbol{f}_{r\leftarrow t}$ are utilized collaboratively to construct a motion cost volume denoted as $\boldsymbol{\mathcal{V}}_m$. The motion-aware cost volume $\boldsymbol{\mathcal{V}}_m$ and the target frame $\boldsymbol{I}_t$ are used in unison as inputs to the attention-based depth network $\boldsymbol{\theta}_{ad}$ with the objective of predicting the final depth map $\boldsymbol{D}_t$. Subsequently, this predicted depth map $\boldsymbol{D}_t$ is utilized in the construction of the ultimate photometric loss denoted as $\mathcal{L}_p$.

C. Pseudo-Static Reference Frame

To address the depth estimation of moving objects with monocular video, we leverage the estimation of flow $\boldsymbol{f}_{r\leftarrow t}$, transformation matrix $[\boldsymbol{R}|\boldsymbol{t}]_{rt}$, and coarse depth $\boldsymbol{D}_t^c$. Using the image warping relationship between $\boldsymbol{I}_t$ and $\boldsymbol{I}_r$, we can calculate a depth-based flow (static flow) $\boldsymbol{f}_{r\leftarrow t}'$ according to the following expression:

$$\boldsymbol{f}_{r\leftarrow t}' = \boldsymbol{p}_{r\leftarrow t}' - \boldsymbol{p}_t, \quad \boldsymbol{p}_{r\leftarrow t}' = \frac{1}{\boldsymbol{D}_r}(\boldsymbol{K}_r)(\boldsymbol{P}_t) \tag{2}$$

where $\boldsymbol{p}_t$ are pixels in the frame $\boldsymbol{I}_t$; $\boldsymbol{P}_t$ are 3D points corresponding to $\boldsymbol{p}_t$ in the frame $\boldsymbol{I}_t$; $\boldsymbol{K}_r$&$\boldsymbol{K}_t$ are intrinsic matrix for frames $\boldsymbol{I}_r$&$\boldsymbol{I}_t$. In static scenarios, the static flow $\boldsymbol{f}_{r\leftarrow t}'$ aligns with the real optical flow $\boldsymbol{f}_{r\leftarrow t}$. As illustrated in Fig. 3, in scenarios involving moving objects, the real optical flow $\boldsymbol{f}_{r\leftarrow t}$ can be decomposed into static optical flow $\boldsymbol{f}_{r\leftarrow t}'$ and dynamic optical flow $\boldsymbol{f}_{r\leftarrow t}^d$. Then the dynamic flow can be calculated as $\boldsymbol{f}_{r\leftarrow t}^d = \boldsymbol{f}_{r\leftarrow t} - \boldsymbol{f}_{r\leftarrow t}'$.

The motion mask could be generated as following:

$$\boldsymbol{\mathcal{M}}_m^r = ||\boldsymbol{f}_{r\leftarrow t} - \boldsymbol{f}_{r\leftarrow t}'||_2^2 > \epsilon, \tag{3}$$

where ϵ is the threshold for distinguishing moving parts. The motion mask in the target mask can be used to generate the motion mask $\boldsymbol{\mathcal{M}}_m^r$ in the reference mask, utilizing $\boldsymbol{f}_{r\leftarrow t}$. With $\boldsymbol{\mathcal{M}}_m^r$, it becomes possible to reconstruct a pseudo-static reference image $\hat{\boldsymbol{I}}_r$ by eliminating the impact of dynamic objects. This process can be formulated as follows:

$$\hat{\boldsymbol{I}}_r = \begin{cases} \boldsymbol{I}_r & \boldsymbol{p}_i \notin \boldsymbol{\mathcal{M}}_m^r; \\ \phi'(\boldsymbol{I}_t, \boldsymbol{D}_t^c, \boldsymbol{K}_t, \boldsymbol{K}_r, [\boldsymbol{R}|\boldsymbol{t}]_{rt}) & \boldsymbol{p}_i \in \boldsymbol{\mathcal{M}}_m^r. \end{cases} \tag{4}$$

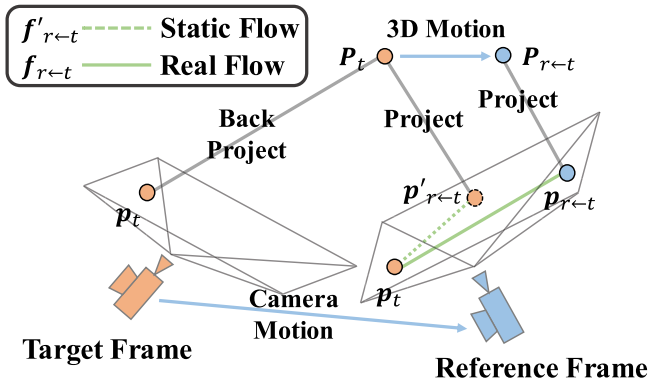

Fig. 3. Relationship between optical flow and depth in the dynamic scenario. Given a point $\boldsymbol{P}_t$ corresponding to pixel $\boldsymbol{p}_t$ in the target frame, it moves to $\boldsymbol{P}_{r\leftarrow t}$ corresponding to pixel $\boldsymbol{p}_{r\leftarrow t}$ in the reference frame. The camera motion causes the pixel to move from $\boldsymbol{p}_t$ to $\boldsymbol{p}'_{r\leftarrow t}$. Both camera motion and 3D object motion contribute to the pixel movement from $\boldsymbol{p}_t$ to $\boldsymbol{p}_{r\leftarrow t}$.

It is important to note that $\phi'(.)$ differs from $\phi(.)$, as $\phi(.)$ is utilized to generate current frame information via sampling based on pixel correspondence between the current frame and the adjacent frame, whereas $\phi'(.)$ generates adjacent frame by shifting forward pixels of the current frame.

D. Motion-Aware Cost Volume

In constructing the motion-aware cost volume $\boldsymbol{\mathcal{V}}_m$ for target frame $\boldsymbol{I}_t$, a set of parallel planes that are perpendicular to the optic axis of $\boldsymbol{I}_t$ is defined, based on the depth assumptions $\boldsymbol{\mathcal{D}} = \{d_k; k = 1, \ldots, M\}$, where M represents the number of planes. The feature extractor $\boldsymbol{\theta}_{fe}$ is utilized to generate the feature maps $\boldsymbol{\mathcal{F}}_t$ and $\hat{\boldsymbol{\mathcal{F}}}_r$ of $\boldsymbol{I}_t$ and $\alpha\hat{\boldsymbol{I}}_r + (1-\alpha)\boldsymbol{I}_r$, respectively. α denotes the hyper-parameters used to balance the influence of pseudo-static reference frame $\hat{\boldsymbol{I}}_r$. With the aid of $\boldsymbol{\mathcal{D}}$ and $[\boldsymbol{R}|\boldsymbol{t}]_{rt}$, a set of wrapped feature maps $\{\hat{\boldsymbol{\mathcal{F}}}^{d_k}_{t\leftarrow r}; d_k \in \boldsymbol{\mathcal{D}}\}$ are generated by warping $\hat{\boldsymbol{\mathcal{F}}}_r$. Considering $\boldsymbol{\mathcal{V}}_m = \{\boldsymbol{\mathcal{P}}_k, k = 1, \ldots, M\}$, it could be written as:

$$\boldsymbol{\mathcal{P}}_k = |\hat{\boldsymbol{\mathcal{F}}}^{d_k}_{t\leftarrow r} - \boldsymbol{\mathcal{F}}_t|. \tag{5}$$

Instead of directly taking the volume $\boldsymbol{\mathcal{V}}_m$ into the attention-based depth network, we propose to construct the final feature volume $\boldsymbol{\mathcal{V}}_f$ by combining the $\boldsymbol{\mathcal{V}}_m$ with the $\boldsymbol{\mathcal{F}}_t$ though channel attention. The process of constructing the final feature volume $\boldsymbol{\mathcal{V}}_f$ is presented in Fig. 2.

E. Attention-Based Depth Network

The High-Resolution Network (HRNet) [17] is well-regarded for its ability to preserve a high level of detail in input images. The HRNet is composed of multiple branches, denoted as B, each generating $S = b, \ldots, 4$ features $\boldsymbol{f}^b_s$ with resolutions of $(\frac{H}{2^{b-1}}, \frac{W}{2^{b-1}})$. However, instead of exclusively utilizing the feature map from the last stage of each branch for disparity map prediction, we leverage the attention mechanism to integrate feature maps from the current branch's various stages and feature maps from deeper branches. To achieve fusion across both the channel and spatial dimensions, we introduce an innovative Depth Attention Network that leverages non-local attention $\mathcal{NA}$ and channel attention $\mathcal{CA}$. Noting that the j^{th} branch of depth decoder takes the $(B-j)^{th}$ branch as input. This fusion process can be expressed as:

$$\boldsymbol{x}^j = \begin{cases} \mathcal{CA}(\boldsymbol{f}^B_B, \quad [\boldsymbol{f}^{B-j-1}_s]^{s=B}_{s=B-j-1}) & j == 0; \\ \mathcal{CA}(\mathcal{NA}(\boldsymbol{x}^{j-1}), \quad [\boldsymbol{f}^{B-j-1}_s]^{s=B}_{s=B-j-1}) & j > 0. \end{cases} \tag{6}$$

F. Loss Function

To summarize, there are three loss functions used in updating Manydepth2's weights. The final loss function $\mathcal{L}$ could be written as: $\mathcal{L} = \mathcal{L}_p + \mathcal{L}_s + \mathcal{L}_c$. The photometric loss $\mathcal{L}_p$ consists of L_1 norm and SSIM regularization:

$$\mathcal{L}_p = a\sum_p \boldsymbol{\mathcal{M}}_o \odot |\boldsymbol{I}_t - \boldsymbol{I}'_t| + b\frac{1 - \text{SSIM}(\boldsymbol{I}_t, \boldsymbol{I}'_t)}{2}, \tag{7}$$

where $\boldsymbol{\mathcal{M}}_o$ is the auto mask introduced in [18]. $\mathcal{L}_s$ is the smooth loss, which could be written as:

$$\mathcal{L}_s = |\partial_x \boldsymbol{D}_t| e^{-|\partial_x \boldsymbol{I}_t|} + |\partial_y \boldsymbol{D}_t| e^{-|\partial_y \boldsymbol{I}_t|}. \tag{8}$$

And $\mathcal{L}_c$ is the consistency loss introduced by [8], to preserve the consistency between prior monocular depth and final multi-frame depth, which could be written as:

$$\mathcal{L}_c = \sum \boldsymbol{\mathcal{M}} \odot |\boldsymbol{D}^c_t - \boldsymbol{D}_t|, \tag{9}$$

where $\boldsymbol{\mathcal{M}}$ is the mask introduced in [8].

IV. Experiments

The main tests and evaluations for our study were carried out using two primary datasets: KITTI-2015 [39] and Cityscapes [40]. In our experiments, the ResNet18 [41] architecture was utilized for the pose network. The HRNet16 was used as the backbone for the attention-based depth network. The pre-trained Gmflow [42] on the Sintel dataset [43] are used as optic flow network. For testing, we set the minimum and maximum depth values to 0.1 m and 80 m, as in [3], [44]. In our experiment, the scale factor is computed using the median value of the ground-truth image as in [9].

A. Monocular Depth Estimation

KITTI-2015: The Table I presents a comparison between the performance of Manydepth2 and state-of-the-art algorithms. Overall, Manydepth2 exhibits superior performance compared to other methods, irrespective of whether monocular or multi-view images are used. In particular, Manydepth2 demonstrates a significant performance advantage over Manydepth, which also employs multiview images and a cost volume structure to estimate depth. Manydepth2 outperforms Manydepth by a substantial margin of 7.2% in terms of Abs Rel. In comparison to Dynamicdepth, which employs semantic motion segmentation maps to leverage high-level computer vision information for precise depth estimation, Manydepth2 achieves a superior performance of 5.3% in terms of the Abs Rel, without the inclusion of any domain-specific information. Quantitative results are shown in Fig. 4. We also present the results for Manydepth2-NF,

TABLE I
SELF-SUPERVISED DEPTH ESTIMATION RESULTS ON THE KITTI 2015

	T	Method	M	S	F	W x H	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
KITTI 2015	Stereo	EPC++ [19]	1			1024 x 320	0.128	0.935	5.011	0.209	0.831	0.945	0.979
		FeatDepth* [20]	1			1024 x 320	0.099	0.697	4.427	0.184	0.889	0.963	0.982
		DFR [21]	1			608 x 160	0.135	1.132	5.585	0.229	0.820	0.933	0.971
	Monocular	Monodepth2* [3]	1			1024 x 320	0.115	0.882	4.701	0.190	0.879	0.961	0.982
		FeatDepth* [20]	1			1024 x 320	0.104	0.729	4.481	0.179	0.893	0.965	0.984
		DevNet* [22]	1			1024 x 320	0.103	0.713	4.459	0.177	0.890	0.965	0.982
		Ranjan [23]	1		•	832 x 256	0.148	1.149	5.464	0.226	0.815	0.935	0.973
		EPC++ [19]	1			832 x 256	0.141	1.029	5.350	0.216	0.816	0.941	0.976
		Guizilini [24]	1	•		640 x 192	0.102	0.698	4.381	0.178	0.896	0.964	0.984
		Monodepth2* [3]	1			640 x 192	0.115	0.903	4.863	0.193	0.877	0.959	0.981
		Patil [11]	2	•		640 x 192	0.111	0.821	4.650	0.187	0.883	0.961	0.982
		Li [25]	2	•		640 x 192	0.102	0.703	4.348	0.175	0.895	0.966	0.984
		ManyDepth* [8]	2			640 x 192	0.098	0.770	4.459	0.176	0.900	0.965	0.983
		Lee [26]	2			640 x 192	0.096	0.644	4.230	0.172	0.903	0.968	0.985
		AQUANet [27]	2			640 x 192	0.105	0.621	4.227	0.179	0.889	0.964	0.984
		DCPI-Depth [28]	2		•	640 x 192	0.095	0.662	4.274	0.170	0.902	0.967	0.985
		DynamicDepth* [15]	2	•		640 x 192	0.096	0.720	4.458	0.175	0.897	0.964	0.984
		Dynamo-Depth* [29]	2			640 x 192	0.112	0.758	4.505	0.183	0.873	0.959	0.984
		DualRefine* [30]	2			640 x 192	0.105	0.787	4.544	0.183	0.891	0.964	0.983
		TC-Depth* [31]	2			640 x 192	0.103	0.746	4.483	0.178	0.894	0.965	0.983
		MOVEDepth* [32]	2			640 x 192	0.097	0.785	4.512	0.177	0.899	0.963	0.982
		DS-Depth [33]	2		•	640 x 192	0.095	0.698	4.329	0.173	0.905	0.966	0.984
		Abdein [34]	2		•	640 x 192	0.095	0.618	3.940	0.168	0.904	0.969	0.988
		ManyDepth2-NF	2			640 x 192	0.094	0.676	4.246	0.170	0.909	0.968	0.985
		Manydepth2	2		•	640 x 192	0.091	0.649	4.232	0.170	0.909	0.968	0.984

Metrics denoted by Color have a lower optimal value, while metrics denoted by Color have a higher optimal value. The optimal results in each subsection are denoted in bold. M: Multi-frames. (1 denotes a single frame, 2 denotes the use of two frames during test time, and so on.) S: Motion Segmentation. F: Optical Flow. Method* denotes the reported result after re-implementing the provided code within the original framework, which may slightly differ from the originally reported numbers. Manydepth2-NF refers to Manydepth2 without optical flow.

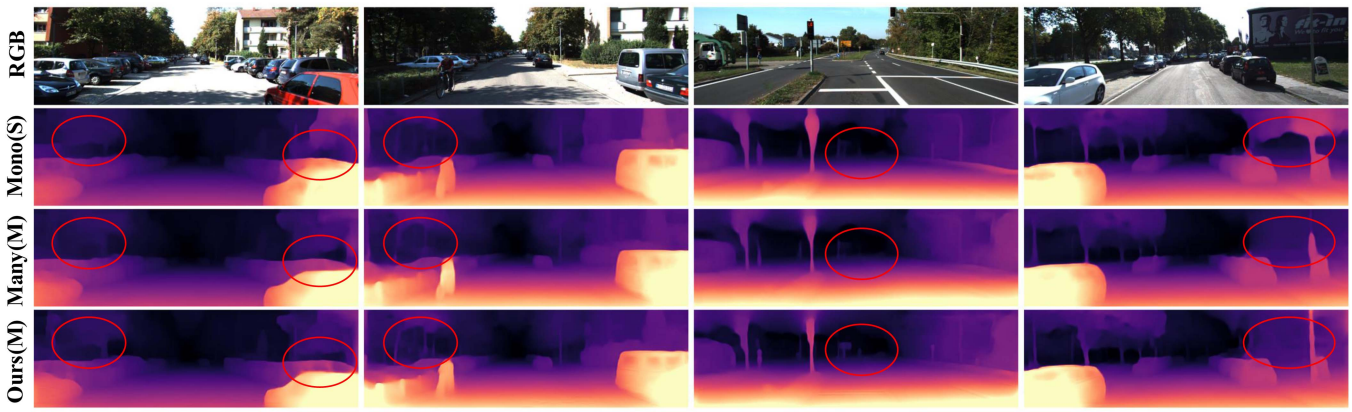

Fig. 4. Qualitative results on KITTI Dataset. The initial row displays the RGB images of target frames where depth information has been estimated. The subsequent rows, specifically the second, third, and fourth, showcase depth maps generated by Monodepth2 trained with stereo techniques, ManyDepth, and Manydepth2, respectively.

a version of Manydepth2 that does not use optical flow or the motion-guided cost volume. We found that our attention-based network outperformed Manydepth by 4.1%.

Cityscapes: In contrast to the KITTI dataset, the Cityscapes dataset features a higher percentage of dynamic objects. During the training phase on the Cityscapes dataset, the input resolution of Manydepth2 is set to 128×416. The lower section of Table II indicates a substantial performance advantage of Manydepth2 over its competing methods. Remarkably, Manydepth2 exhibits superior performance over ManyDepth, with a margin of approximately 15.0% in the absolute relative error metric. Through the visualization of error maps generated by computing the differences between predicted depth maps and their corresponding ground truth depth maps, Fig. 1 facilitates a qualitative analysis of depth predictions. These observations indicate that Manydepth2 offers enhanced accuracy for depth estimation of both dynamic foreground and static background. Additionally, the incorporation of a pre-trained optical flow estimation model with low computational requirements has resulted in a minimal increase in parameter size and running time for Manydepth2. Quantitative results are shown in Fig. 1. Table III showcases the depth errors within dynamic object regions by utilizing the ground truth motion segmentation maps provided in [16], which corresponds to the qualitative results in Fig. 1.

B. Odometry Estimation

To evaluate odometry estimation results, we follow the split used in [9], [21]. Specifically, we trained our model using Seq. 00-08 from the KITTI odometry dataset and test methods on

TABLE II
SELF-SUPERVISED DEPTH ESTIMATION RESULTS ON THE CITYSCAPES

	T	Method	M	S	F	W x H	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Cityscapes	Monocular	Lee [35]	1			832 x 256	0.116	1.213	6.695	0.186	0.852	0.951	0.982
		InstaDM [16]	1	•		832 x 256	0.111	1.158	6.437	0.182	0.868	0.961	0.983
		Pilzer [36]	1			512 x 256	0.240	4.264	8.049	0.334	0.710	0.871	0.937
		Monodepth2* [3]	1			416 x 128	0.129	1.569	6.876	0.187	0.849	0.957	0.983
		Videos in the Wild [37]	1			416 x 128	0.127	1.330	6.960	0.195	0.830	0.947	0.981
		Li [38]	1			416 x 128	0.119	1.290	6.980	0.190	0.846	0.952	0.982
		Struct2Depth [10]	2			416 x 128	0.151	2.492	7.024	0.202	0.826	0.937	0.972
		ManyDepth* [8]	2			416 x 128	0.114	1.193	6.223	0.170	0.875	0.967	0.989
		DynamicDepth* [15]	2	•	•	416 x 128	0.103	1.000	5.867	0.157	0.895	0.974	0.991
		Manydepth2	2		•	416 x 128	0.097	0.792	5.827	0.154	0.903	0.975	0.993

Metrics denoted by Color have a lower optimal value, while metrics denoted by Color have a higher optimal value. The optimal results in each subsection are denoted in bold. Our method outperforms all prior approaches in most metrics across all subsections, irrespective of whether the baselines utilize multiple frames during testing. M: Multi-frames. S: Motion Segmentation. F: Optical Flow.

TABLE III
RESULTS OF DYNAMIC FOREGROUND ON CITYSCAPES

Type	Method	M	WxH	Abs Rel	Sq Rel	RMSE	RMSE log
S	InstaDM [16]	1	832 x 256	0.139	1.698	5.760	0.181
w/o S	Monodepth2 [3]	1	416 x 128	0.159	1.937	6.363	0.201
	ManyDepth [8]	2	416 x 128	0.169	2.175	6.634	0.218
	Manydepth2-NF	2	416 x 128	0.143	1.483	5.123	0.179
	Manydepth2	2	416 x 128	0.123	1.260	4.519	0.144

We evaluate depth prediction errors for dynamic objects such as vehicles, people, and bikes on the Cityscapes dataset [40]. The best results are highlighted in bold.

TABLE IV
VISUAL ODOMETRY RESULTS ON SEQ. 9 & SEQ. 10 OF THE KITTI ODOMETRY

Method	Tr	R	Tr	R
DFR [21]	11.93	3.91	12.45	3.46
NeuralBundler [45]	8.10	2.81	12.90	3.71
ManyDepth [8]	8.08	1.97	9.86	3.42
Manydepth2	7.01	1.76	7.29	2.65

The average drift in root mean square error for translation (Tr) and rotation (R) is reported.

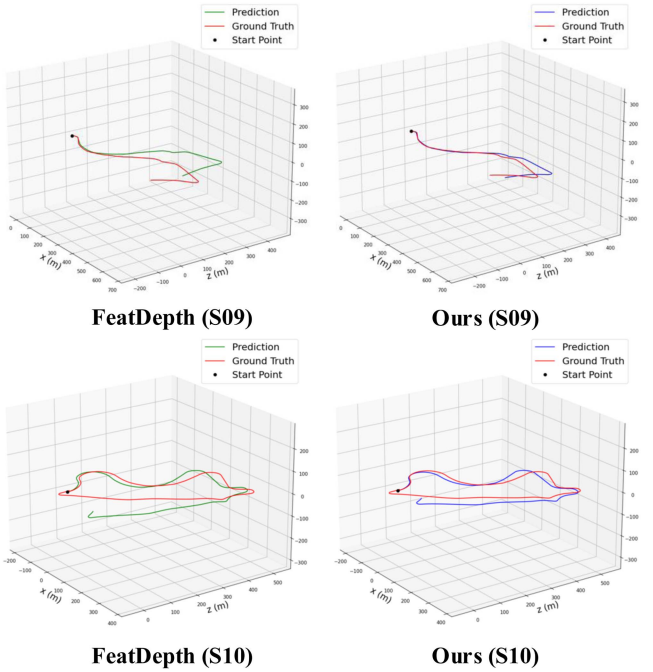

Fig. 5. Trajectory visualization on Seq. 09. and Seq. 10. The ground truth trajectories are represented by red lines. Trajectories generated by the FeatDepth [20] results are indicated by green lines, while our method's trajectories are depicted in blue lines.

Seq. 09-10. Table IV shows the results, where the translational and rotational motion were evaluated by using the root mean square error (RMSE). Manydepth2, which uses less computational resources, outperforms other learning-based techniques that use more computational resources, such as FeatDepth. The results indicate that Manydepth2 achieves a significant reduction in RMSE by approximately 24.1% for translational motion and 22.5% for rotational motion, as compared to ManyDepth, on Seq. 10. In support of this result, we present a visual depiction of the trajectory of both methods in Fig. 5, which shows that Manydepth2's trajectory exhibits considerably less drift than that of FeatDepth. The superior performance of Manydepth2 in precise pose estimation can be attributed to its use of both depth-based photometric loss and flow-based ego-motion loss, as well as the implementation of a motion-guided mask that effectively filters out dynamic foreground outliers. This combination of techniques enables Manydepth2 to achieve more reliable pose estimates compared to others.

C. Abaltion Study

In our research, we meticulously carry out ablation studies using the Cityscapes dataset. These studies are designed to dissect and showcase how effective our proposed modules are, allowing us to validate the specific design choices we've made.

TABLE V
ABLATION STUDY RESULTS OF DEPTH ACCURACY ON THE KITTI 2015

Ablation	Abs Rel	Sq Rel	RMSE
Final Prior Depth	0.105	0.768	4.536
w/o Attention for Volume	0.100	0.709	4.298
Manydepth2-NF	0.094	0.676	4.246
Manydepth2	0.091	0.649	4.232

The proposed Manydepth2 model was evaluated by modifying its components.

The detailed outcomes of these experiments are showcased in Table V. Within this table, we provide insights from experiments conducted under diverse conditions and configurations. Specifically, we assess Manydepth2's performance with variations such as the Prior Depth Net, Manydepth2 configurations without Motion-Guided Volume and without Attention for Final Volume. A key observation from Table V is the pronounced

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

positive impact that integrating the Attention-Based HRNet has on enhancing Manydepth2's performance. Furthermore, our findings underscore that the incorporation of both the Motion-Guided Volume and Attention for the Final Volume amplifies the capabilities of Manydepth2 even further, leading to improved outcomes.

V. Conclusion

We introduce Manydepth2, a self-supervised multi-frame monocular depth prediction model that leverages optical flow-depth geometry. Key features include a static reference frame, motion-guided cost volume, and an attention-enhanced high-resolution neural network. Benchmark results on KITTI-2015 and Cityscapes show accurate depth estimation, distinguishing dynamic objects from static backgrounds, all while maintaining efficiency on a single RTX 3090. Manydepth2 enhances existing methods by incorporating dynamic insights, significantly improving accuracy.

References

[1] K. Zhou et al., "Dynpoint: Dynamic neural point for view synthesis," in *Proc. Conf. Neural Inf. Process. Syst.*, 2023, pp. 69532–69545.

[2] S. Li, J. Shi, W. Song, A. Hao, and H. Qin, "Hierarchical object relationship constrained monocular depth estimation," *Pattern Recognit.*, vol. 120, 2021, Art. no. 108116.

[3] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3828–3838.

[4] K. Zhou, L. Hong, E. Xie, Y. Yang, Z. Li, and W. Zhang, "Splatmesh: Interactive 3D segmentation and editing using mesh-based Gaussian splatting," 2023, *arXiv:2312.15856*.

[5] K. Zhou, "Neural surface reconstruction from sparse views using Epipolar geometry," 2024, *arXiv:2406.04301*.

[6] J.-W. Bian et al., "Unsupervised scale-consistent depth learning from video," *Int. J. Comput. Vis.*, vol. 129, no. 9, pp. 2548–2564, 2021.

[7] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 767–783.

[8] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1164–1174.

[9] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1851–1858.

[10] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8001–8008.

[11] V. Patil, W. Van Gansbeke, D. Dai, and L. Van Gool, "Don't forget the past: Recurrent depth estimation from monocular video," *IEEE Robot. Automat. Lett.*, vol. 5, no. 4, pp. 6813–6820, Oct. 2020.

[12] F. Wimbauer, N. Yang, L. von Stumberg, N. Zeller, and D. Cremers, "MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6112–6122.

[13] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2495–2504.

[14] M. Klingner, J.-A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, "Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., Springer, 2020, pp. 582–600.

[15] Z. Feng, L. Yang, L. Jing, H. Wang, Y. Tian, and B. Li, "Disentangling object motion and occlusion for unsupervised multi-frame monocular depth," in *Proc. 17th Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Springer, 2022, pp. 228–244.

[16] S. Lee, S. Im, S. Lin, and I. S. Kweon, "Learning monocular depth in dynamic scenes via instance-aware projection consistency," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1863–1872.

[17] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[18] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 270–279.

[19] C. Luo et al., "Every pixel counts : Joint learning of geometry and motion with 3D holistic understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2624–2641, Oct. 2020.

[20] C. Shu, K. Yu, Z. Duan, and K. Yang, "Feature-metric loss for self-supervised learning of depth and egomotion," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 572–588.

[21] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 340–349.

[22] K. Zhou et al., "DevNet: Self-supervised monocular depth learning via density volume construction," in *Proc. 17th Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, Springer, 2022, pp. 125–142.

[23] A. Ranjan et al., "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12240–12249.

[24] V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon, "Semantically-guided representation learning for self-supervised monocular depth," 2020, *arXiv:2002.12319*.

[25] R. Li et al., "Learning depth via leveraging semantics: Self-supervised monocular depth estimation with both implicit and explicit semantic guidance," *Pattern Recognit.*, vol. 137, 2023, Art. no. 109297.

[26] S. Lee, W. Im, and S.-E. Yoon, "Multi-resolution distillation for self-supervised monocular depth estimation," *Pattern Recognit. Lett.*, vol. 176, pp. 215–222, 2023.

[27] J. L. G. Bello, J. Moon, and M. Kim, "Self-supervised monocular depth estimation with positional shift depth variance and adaptive disparity quantization," *IEEE Trans. Image Process.*, vol. 33, pp. 2074–2089, 2024.

[28] M. Zhang, Y. Feng, Q. Chen, and R. Fan, "DCPI-Depth: Explicitly infusing dense correspondence prior to unsupervised monocular depth estimation," 2024, *arXiv:2405.16960*.

[29] Y. Sun and B. Hariharan, "Dynamo-Depth: Fixing unsupervised depth estimation for dynamical scenes," in *Proc. Adv. Neural Inf. Proc. Syst.*, 2023, pp. 54987–55005.

[30] A. Bangunharcana, A. Magd, and K.-S. Kim, "DualRefine: Self-supervised depth and pose estimation through iterative epipolar sampling and refinement toward equilibrium," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 726–738.

[31] P. Ruhkamp, D. Gao, H. Chen, N. Navab, and B. Busam, "Attention meets geometry: Geometry guided spatial-temporal attention for consistent self-supervised monocular depth estimation," in *Proc. 2021 Int. Conf. 3D Vis.*, IEEE, 2021, pp. 837–847.

[32] X. Wang et al., "Crafting monocular cues and velocity guidance for self-supervised multi-frame depth learning," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 2689–2697.

[33] X. Miao et al., "DS-Depth: Dynamic and static depth estimation via a fusion cost volume," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2564–2576, Apr. 2024.

[34] R. Abdein, W. Li, C. Li, X. Zheng, and R. Yadav, "Self-supervised uncertainty-guided refinement for robust joint optical flow and depth estimation," in *Proc. 2025 IEEE Int. Conf. Acoust. Speech Signal Process.*, IEEE, 2025, pp. 1–5.

[35] S. Lee, F. Rameau, F. Pan, and I. S. Kweon, "Attentive and contrastive learning for joint depth and motion field estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4862–4871.

[36] A. Pilzer, D. Xu, M. Puscas, E. Ricci, and N. Sebe, "Unsupervised adversarial depth estimation using cycled generative networks," in *Proc. 2018 Int. Conf. 3D Vis.*, IEEE, 2018, pp. 587–595.

[37] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8977–8986.

[38] H. Li, A. Gordon, H. Zhao, V. Casser, and A. Angelova, "Unsupervised monocular depth learning in dynamic scenes," in *Proc. Conf. Robot Learn.*, PMLR, 2021, pp. 1908–1917.

[39] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The kitti vision benchmark suite," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[40] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[42] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "GMFlow: Learning optical flow via global matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8121–8130.

[43] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. 12th Eur. Conf. Comput. Vis.*, Florence, Italy, Springer, 2012, pp. 611–625.

[44] H. Jin, P. Favaro, and S. Soatto, "Real-time feature tracking and outlier rejection with changes in illumination," in *Proc. Int. Conf. Comput. Vis.*, 2001, pp. 684–689.

[45] Y. Li, Y. Ushiku, and T. Harada, "Pose graph optimization for unsupervised monocular visual odometry," in *Proc. 2019 Int. Conf. Robot. Automat.*, IEEE, 2019, pp. 5439–5445.