ReloDepth: Reprojection Loss Filtering for Self-Supervised Depth Estimation in Dynamic Environments

Siyu Chen[1], Hong Liu[1,*], Wenhao Li[2], Ying Zhu[1], Jianbing Wu[1], and Guoquan Wang[1]
[1]State Key Laboratory of General Artificial Intelligence, Peking University Shenzhen Graduate School, Shenzhen, China
[2]Nanyang Technological University, Singapore
Email: {csyunling, YingZhu, guoquanwang}@stu.pku.edu.cn, hongliu@pku.edu.cn,
wenhao.li@ntu.edu.sg, kimbingng@gmail.com

***Abstract*—Depth estimation plays a pivotal role in robotics applications, where self-supervised approaches have emerged as promising solutions by leveraging abundant unlabelled real-world data. While existing methods achieve notable performance under static scene assumptions, their effectiveness significantly degrades in dynamic environments due to moving objects. To address this issue, we present ReloDepth, a novel method for self-supervised depth estimation in dynamic scenes. It tackles the challenge of dynamic objects from two key perspectives. First, within the self-supervised learning framework, we introduce a reprojection loss mask to explicitly localize regions likely contaminated by dynamic objects, thereby mitigating their corrupting influence on the reprojection-based supervisory signal at the loss level. Second, for multi-frame depth estimation, we propose a photometric gated cost volume strategy to mask redundant regions prior to cost volume, improving depth estimation robustness by steering the attention of the network toward photometrically informative regions. Furthermore, we propose a spectral entropy uncertainty module that incorporates spectral entropy to guide uncertainty estimation during depth fusion, effectively addressing issues arising from cost volume computation in dynamic environments. Extensive experiments on KITTI and Cityscapes datasets demonstrate that the proposed method consistently outperforms existing self-supervised depth estimation baselines.**

***Index Terms*—Self-supervised Learning, Depth Estimation, Reprojection Loss Filtering**

I. INTRODUCTION

Depth information is crucial for various real-world applications, such as autonomous vehicles [1], robot navigation [2], and human-robot interaction [3]. Self-supervised depth estimation [4] has emerged as a promising method due to its ability to predict pixel-level depth maps from images without requiring annotated data. Single-frame methods predict depth from a single target image, employing photometric reprojection of adjacent frames for supervision [5]. Multi-frame methods [6], [7], on the other hand, utilize consecutive frames during inference, integrating richer information and achieving higher performance. However, both single-frame and multi-frame self-supervised methods are typically built upon the assumption of static scenes, presenting challenges when dynamic objects are encountered in real-world environments. This stems from two primary factors:

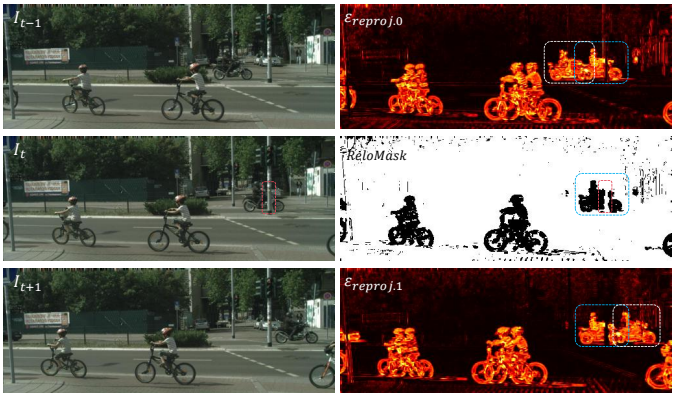

Fig. 1. Visualization of reprojection loss mask and reprojection error maps. The left column shows the input frames I_t, I_{t-1}, I_{t+1}. The first and third rows of the right column visualize the reprojection error maps ($\mathcal{E}_{\text{reproj.0/1}}$), which are obtained by projecting the source images (I_{t-1}, I_{t+1}) back to the target image (I_t). Darker regions in the error maps correspond to lower errors, while brighter areas indicate regions of higher error. The second row presents our proposed reprojection loss mask (ReloMask), generated from the combination of the two reprojection error maps, highlighting regions that potentially exhibit dynamic errors.

(i) The prevailing framework of self-supervised depth estimation [4], [8] relies on the reprojection loss to model geometric constraints between consecutive frames under the assumption of photometric consistency. However, this assumption breaks down in the presence of dynamic objects, as shown in Fig. 1, where a person riding a bike leads to errors during the view synthesis phase, significantly degrading the accuracy of the generated depth maps. To mitigate the impact of dynamic objects, several existing approaches have proposed solutions. SfMLearner [5] introduces an explainability mask, though it merely adds regularization constraints. Monodepth2 [9] employs a minimum reprojection loss strategy, while SC-DepthV3 [4] incorporates pseudo-depth information for optimization.

(ii) In multi-frame depth estimation, the construction of cost

*Corresponding author: Hong Liu (hongliu@pku.edu.cn). This research was supported by the National Key R&D Program of China (No. 2024YFB4700271) and Guangdong S&T Program (No. 2024B0101050002).

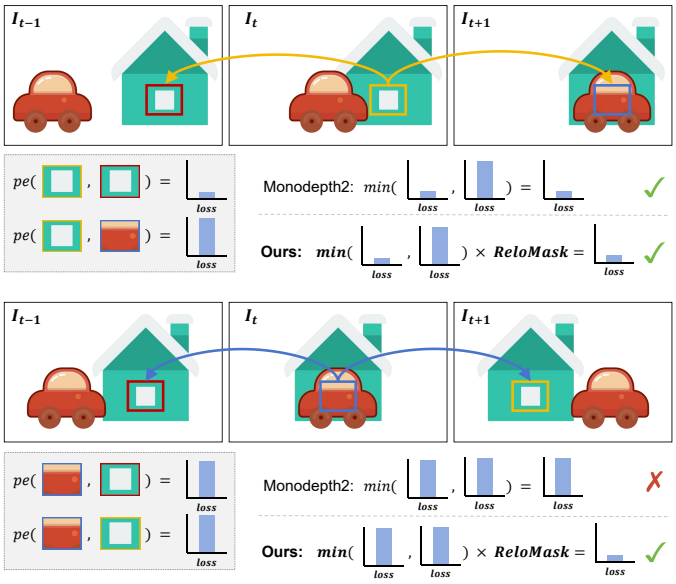

Fig. 2. **Comparison of our ReloDepth with Monodepth2.** The upper part of the figure illustrates the typical occlusion scenario addressed by Monodepth2. Both our method and Monodepth2 can handle this situation. However, the lower part of the figure depicts a case where Monodepth2 fails to resolve, as it relies on the minimum of two reprojection losses. In contrast, our method effectively addresses this issue by performing an additional calculation of a reprojection loss mask (ReloMask) to correct the loss.

volumes does not account for dynamic objects and occlusions, introducing additional errors and making multi-frame depth estimation more susceptible to challenges posed by dynamic scenes. To alleviate this, some studies [6], [7] employs teacher-student distillation, where a single-frame depth network guides the multi-frame depth estimation network to correct feature matching errors caused by cost volumes. Other methods involve semantic segmentation to identify dynamic objects and adjust their matching [7], [10] or use optical flow estimation [11] to alleviate the impact of dynamic objects.

In this paper, we propose **ReloDepth** (**Re**projection **Lo**ss Filtering for Self-Supervised **Depth** Estimation in Dynamic Environments), a novel method for addressing the above challenging problems in dynamic scenes. Specifically, Within the general self-supervised framework, we propose to mitigate this issue through a reprojection loss mask that automatically identifies regions potentially corrupted by dynamic objects, effectively reducing their influence at the loss computation level while introducing zero additional inference overhead. As illustrated in Fig. 1, given three consecutive frames, we compute two sets of reprojection losses. If both sets exhibit high losses in certain regions (blue box), it suggests that these regions are affected by dynamic objects, enabling us to mask them out. Our method builds upon the minimum reprojection loss proposed in Monodepth2 [9] and offers a more robust solution for handling occlusions and dynamic objects. Monodepth2 addresses occlusions, as illustrated in the upper part of Fig. 2, where one of the adjacent frames is occluded. In this case, taking the minimum of the reprojection losses from two frames helps correct the issue. However, Monodepth2 fails to handle the scenario depicted in the lower part of Fig. 2, where a region in the current frame is occluded by a dynamic object, while the corresponding regions in both the previous and subsequent frames remain unobstructed. Our method effectively addresses this case by identifying dynamic regions and correcting the reprojection loss, whereas Monodepth2 struggles with occlusions caused by dynamic objects due to its reliance on the minimum of two high-loss areas.

From the perspective of depth estimation in DepthNet, we focus on multi-frame depth estimation. This approach constructs a cost volume through feature matching to regress depth values, which makes it more susceptible to issues caused by dynamic objects compared to single-frame depth estimation. To overcome this limitation, we propose photometric gated cost volume that intelligently filters out similarity-redundant regions by analyzing inter-frame feature correlations, effectively directing the model's attention to distinctive feature-rich areas while improving depth estimation accuracy. Furthermore, we integrate spectral entropy after the cost volume to capture richer information for uncertainty estimation and depth fusion, effectively addressing the dynamic challenges introduced by cost volumes. In summary, the contributions of this paper are as follows:

- We propose ReloDepth for self-supervised depth estimation in dynamic environments. It introduces the reprojection loss mask (ReloMask) to address the problem of dynamic objects by masking regions with high reprojection losses, enhancing results without extra computational cost during inference.
- We introduce photometric gated cost volume (PGCV) that eliminates feature redundancy through inter-frame consistency validation before cost aggregation, forcing the model to focus on geometrically discriminative regions; and a spectral entropy uncertainty module (SEUM) that generates frequency-aware confidence maps to optimize depth fusion through uncertainty-guided weighting.
- Our ReloDepth achieves state-of-the-art results on the KITTI and Cityscapes datasets.

II. RELATED WORK

A. Single-Frame Self-Supervised Depth Estimation

Self-supervised monocular depth estimation has gained significant attention for predicting pixel-level depth maps without labeled data, with the foundational framework introduced by Zhou et al. [5] via DepthNet and PoseNet for frame geometric relationship prediction. Monodepth2 [9] addresses occlusion issues by proposing the minimum reprojection loss and employs full-resolution multi-scale sampling to mitigate visual artifacts. Further advancements are made by MonoProb [12], which incorporates an interpretable confidence measure to enhance depth modeling through uncertainty estimation. SQLdepth [13] introduces a self-supervised method utilizing the Self Query Layer to build a self-cost volume, effectively capturing fine-grained scene geometry from a single image. To handle

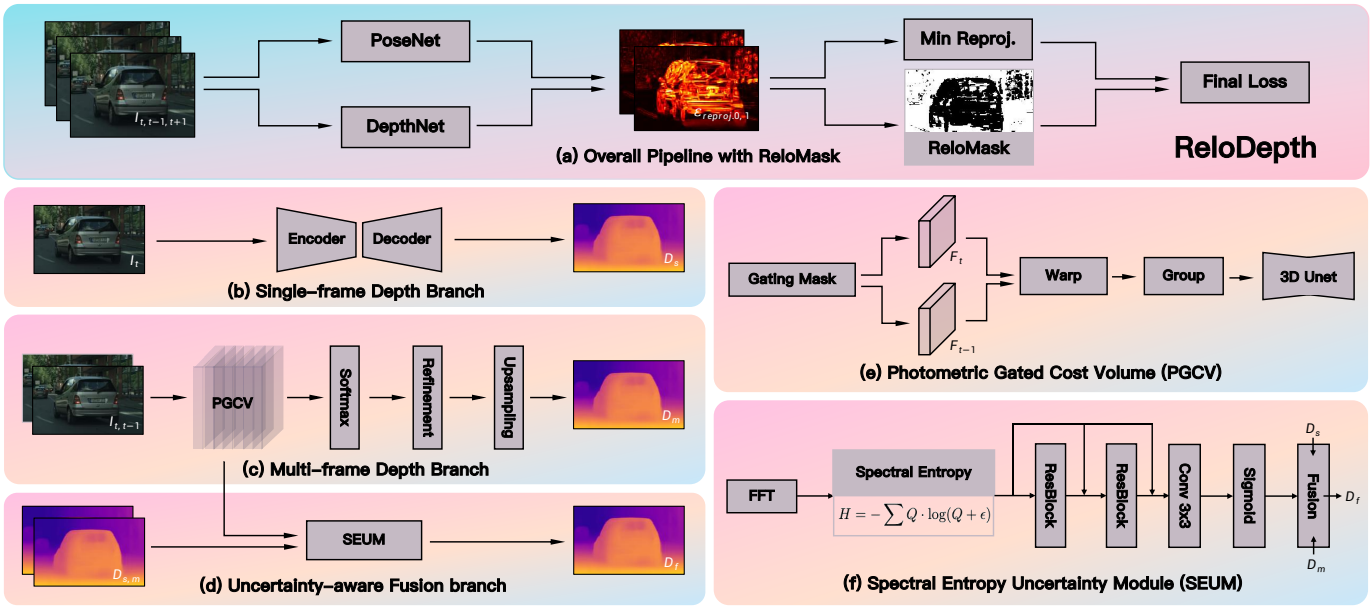

Fig. 3. Overview of the ReloDepth architecture. (a) depicts the integrated self-supervised pipeline featuring the proposed ReloMask module. The DepthNet architecture comprises three specialized branches: the single-frame depth branch (b) producing D_s, the multi-frame depth branch (c) utilizing a photometric gated cost volume (PGCV) for attention-aware regression to yield D_m, and the uncertainty-aware fusion branch (d) generating D_f through spectral entropy-based uncertainty fusion. All outputs undergo supervision via photometric reprojection loss within the self-supervised framework (a). The ReloMask enhances learning robustness through optimized reprojection error weighting.

adverse weather conditions, WeatherDepth [14] introduces a curriculum-based contrastive learning strategy, integrating robust weather adaptation techniques to tackle challenging environments without inducing knowledge forgetfulness.

B. Multi-Frame Self-Supervised Depth Estimation

Compared to single-frame methods, multi-frame depth estimation leverages multiple images for inference, providing richer information. This approach has high practical value in applications such as autonomous driving, making it an increasing focus in recent research. Early approaches [15] to multi-frame depth estimation often employ Recurrent Neural Networks (RNNs) to enhance model performance; however, these methods typically come with high computational costs. ManyDepth [6] introduces an adaptive cost volume feature matching scheme to exploit geometric information between frames using multiple inputs at test time. Building on this, Guizilini et al. [16] propose a transformer architecture for cost volume generation, improving multi-view feature matching. To address dynamic object challenges, DynamicDepth [7] designs an occlusion-aware cost volume to facilitate geometric reasoning in regions occluded by motion. MAL [10] introduces a motion-aware loss and a distillation scheme to correct errors caused by object motion in multi-frame self-supervised depth estimation. DS-Depth [11] employs a dynamic cost volume with residual optical flow to better handle moving objects and occlusions, complemented by a fusion module balancing static and dynamic cost volumes. Moon et al. [17] proposes a coarse-to-fine estimation framework incorporating ground contact priors for handling dynamic objects. However, such approaches often rely on auxiliary modules like semantic segmentation and optical flow, resulting in computationally complex pipelines. In contrast, our method primarily addresses the dynamic object problem from the loss level. Both the proposed reprojection loss mask and spectral entropy uncertainty module do not introduce additional inference costs.

III. METHOD

A. Preliminary

Self-supervised methods use reprojection loss between adjacent frames as supervision for training. Specifically, each training instance includes a reference frame I_t and two temporally adjacent source frames I_s ($s \in \{t-1, t+1\}$). The depth map D_t of the target view is estimated by the depth estimation network θ_{depth}, as expressed in the following equation:

$$D_t = \theta_{\text{depth}}(I_t). \tag{1}$$

Additionally, a pose estimation network is used to estimate the camera pose $T_{t\to s}$ between the target and source views. The self-supervised training signal is derived from the reprojection error between adjacent frames. In particular, the image synthesis from the source to the target view, based on the estimated depth D_t and the camera pose $T_{t\to s}$, can be formulated as:

$$I_{s\to t} = I_s \left\langle \text{Proj}(D_t, T_{t\to s}, K) \right\rangle, \tag{2}$$

where K represents the known camera intrinsics, $\langle\cdot\rangle$ is the sampling operator [18], and $\text{Proj}(\cdot)$ is the coordinate projection operation [5]. After obtaining the reconstructed image, the reprojection loss is used to supervise the training process. The

photometric loss is composed of the L1 loss and SSIM [19], and is defined as:

$$pe(\boldsymbol{I}_a, \boldsymbol{I}_b) = \frac{\alpha}{2}(1 - SSIM(\boldsymbol{I}_a, \boldsymbol{I}_b)) + (1-\alpha)\left\|\boldsymbol{I}_a - \boldsymbol{I}_b\right\|_1. \tag{3}$$

Following the approach in Monodepth2 [9], we adopt the per-pixel minimum reprojection loss, which is defined as:

$$\mathcal{L}_{rep}(\boldsymbol{I}_t, \boldsymbol{I}_{s\to t}) = \min_s pe(\boldsymbol{I}_t, \boldsymbol{I}_{s\to t}). \tag{4}$$

In addition, we employ an edge-aware smoothness loss [20] $\mathcal{L}_s$ to encourage the depth map to be smooth in local regions while preserving image edge information. This ensures that the depth map better aligns with the geometric structure of the real-world scene. The equation is as follows:

$$\mathcal{L}_s = |\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^*| e^{-|\partial_y I_t|}, \tag{5}$$

where d_t^* represents the mean-normalized inverse depth.

B. Overview

Our framework addresses dynamic object challenges through two key innovations, as illustrated in Fig. 3. First, we introduce a reprojection loss mask (Fig. 3 (a), Sec. III-C) that resolves dynamic object interference at the loss level within the self-supervised learning framework. Second, we enhance multi-frame depth estimation with two components: the photometric gated cost volume (Fig. 3(e), Sec. III-D), which selectively eliminates inter-frame redundancy while preserving discriminative features, and the spectral entropy uncertainty module (Fig. 3(f), Sec. III-E), which leverages spectral entropy analysis to improve uncertainty estimation and enable robust depth fusion. The technical details of each component are elaborated in subsequent sections.

C. Reprojection Loss Mask

In self-supervised depth estimation, dynamic objects frequently violate the photometric consistency assumption, leading to erroneous supervisory signals. While Monodepth2 [9] mitigates simple occlusions by selecting the minimum reprojection loss across adjacent frames (as shown in the white boxes in Fig.1), it fails when regions are consistently corrupted in both temporal directions (e.g., the blue boxes in Fig.1). To address this, we propose a robust Filtering Mask that identifies regions where the photometric assumption is invalid. Our core observation is that if a pixel exhibits high reprojection errors across all available source frames, it likely belongs to a dynamic object or a truly occluded area. By explicitly masking out these high-error regions, our method ensures that only reliable supervisory signals are used for optimization. As illustrated in the red box of Fig.1, this mask effectively filters out dynamic artifacts (e.g., moving cyclists) while preserving static structures (e.g., poles) that are correctly aligned. By selecting only reliable supervisory signals, our approach ensures precise optimization without introducing additional inference overhead, maintaining an elegant balance between performance and computational efficiency.

Specifically, given the reprojection error map $\mathcal{E}_{\text{reproj}} \in \mathbb{R}^{B\times C\times H\times W}$, where B denotes the batch size, $C = 2$ represents the two error channels ($\mathcal{E}_{\text{reproj.0}}$ and $\mathcal{E}_{\text{reproj.1}}$ in Fig. 1), and $H \times W$ are the spatial dimensions, we first flatten each channel's spatial dimensions into a vector:

$$\mathcal{E}_c = \text{Flatten}(\mathcal{E}_{\text{reproj}}[:, c, :, :]) \in \mathbb{R}^{B\times(H\cdot W)}, \tag{6}$$

where $c \in \{0, 1\}$ indexes the error channels. The β-th quantile q_c is computed for each channel to identify high-error regions:

$$q_c = Q_\beta(\mathcal{E}_c), \tag{7}$$

where $Q_\beta(\cdot)$ extracts the β-th quantile value. In our implementation, we set $\beta = 0.8$. The threshold β determines the sensitivity of the quantile-based selection: a higher β excludes only regions with the highest losses, while a lower value of β captures a broader range of potential error regions. Using these computed quantiles, we then generate a binary mask for each channel as follows:

$$\mathcal{M}_c = \mathbb{I}(\mathcal{E}_{\text{reproj}}[:, c, :, :] > q_c), \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Finally, the reprojection loss mask is constructed by combining the binary masks from both channels:

$$\mathcal{M}_{relo} = 1 - (\mathcal{M}_1 \wedge \mathcal{M}_2), \tag{9}$$

which results in a mask that highlights regions identified as high-loss in both reprojection criteria simultaneously.

D. Photometric Gated Cost Volume

Building upon the multi-frame depth estimation framework, we propose photometric gated cost volume (PGCV) as shown in Fig. 3(e), an enhanced construction method that addresses the limitations of traditional approaches in complex environments, particularly dynamic scenes. Conventional methods construct cost volumes by warping the source feature map $F_s \in \mathbb{R}^{C\times H'\times W'}$ to the target view coordinates through differentiable warping $\mathcal{W}(\cdot)$ [5]:

$$FV_{s\to t} = \mathcal{W}(F_s, T_{t\to s}, D_{hypoth}), \tag{10}$$

where $T_{t\to s}$ denotes the relative camera pose and $D_{hypoth} \in \mathbb{R}^{N\times H'\times W'}$ contains N discrete depth hypotheses. The warping operation produces a feature volume $FV_{s\to t} \in \mathbb{R}^{N\times C\times H'\times W'}$ with spatial dimensions $H' = H/4$ and $W' = W/4$. The final cost volume CV is obtained by computing the L1 difference between the warped feature volume $FV_{s\to t}$ and the target frame feature F_t. However, in practical applications, adjacent frames frequently contain low-texture regions or areas with minimal inter-frame variation. These homogeneous regions typically provide poor depth cues while introducing noise and redundancy, ultimately reducing the cost volume's discriminative power. Our PGCV addresses this by actively suppressing non-informative regions (e.g., low-texture areas) during cost volume computation through photometric gating.

Photometric Gating Mechanism. Given a reference frame I_t and a source frame I_{t-1}, we compute their photometric

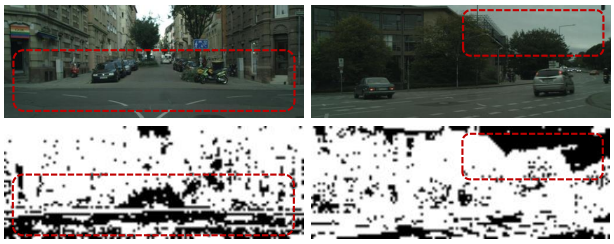

Fig. 4. Photometric Gated Cost Volume (PGCV) visualization. (a) Input images; (b) Our Photometric Gated mask, showing effective suppression of low-texture regions (sky/ground) while preserving structural details.

error $pe(I_t, I_{t-1})$ following Eq. 3. The spatial gating mask is defined using a fixed threshold τ_{pe}:

$$\mathcal{M}(u) = \mathbb{I}[pe(u) > \tau_{pe}] \in \{0,1\}^{H\times W}, \tag{11}$$

where u indexes pixel coordinates, $\mathbb{I}[\cdot]$ is the indicator function, and τ_{pe} is a predefined threshold (set to 0.03 in our experiments). This indicates that regions with similar photometric errors between adjacent frames, typically corresponding to textureless areas, will be filtered out by the gating mechanism. Then the mask is downsampled to match the feature resolution and applied to both reference and source features:

$$\hat{F}_t = F_t \odot \mathcal{M}, \quad \hat{F}_{t-1} = F_{t-1} \odot \mathcal{M}, \tag{12}$$

where $\odot$ indicates element-wise multiplication. This mechanism selectively preserves features in regions with meaningful photometric variation while suppressing ambiguous areas.

Cost Volume Construction. Building upon the gated features $\hat{F}_t$ and $\hat{F}_{t-1}$, we construct the cost volume via group-wise correlation [21]. For each pixel $u = (x, y)^\top$ in the reference frame, depth hypothesis $d \in \{d_1, ..., d_N\}$, and feature channel $c \in \{1, ..., C\}$, we compute the per-channel correlation through differentiable warping:

$$CV_{\text{raw}}(u, d, c) = \hat{F}_t^c(u) \cdot \hat{F}_{t-1}^c(\mathcal{W}(u, T, d)), \tag{13}$$

where $\mathcal{W}(u, T, d)$ denotes the warping operation that projects pixel u from the source to reference frame using camera pose T and depth d. These correlations are then aggregated into a grouped cost volume $CV_{group} \in \mathbb{R}^{N\times G\times H'\times W'}$ by averaging over channel groups:

$$CV_{group}(u, d, g) = \frac{1}{|G_g|} \sum_{c\in G_g} CV_{\text{raw}}(u, d, c), \tag{14}$$

where $G = 16$ groups evenly partition the C channels ($|G_g| = C/G$ per group). The grouped volume is subsequently processed by a 3D UNet $\mathcal{U}_\theta(\cdot)$ to generate the gated cost volume:

$$CV_{gated} = \mathcal{U}_\theta(CV_{group}) \in \mathbb{R}^{N\times H'\times W'}. \tag{15}$$

Depth Regression. The gated cost volume CV_{gated} is normalized into a probability distribution $P_t \in \mathbb{R}^{N\times H'\times W'}$ over depth hypotheses through softmax transformation:

$$P_t = \text{softmax}(CV_{gated}). \tag{16}$$

To achieve sub-bin precision beyond discrete depth bins, we adopt the local-max refinement strategy [22]. The optimal depth bin index $k_{max} = \arg\max_k P_t(k, u, v)$ is first identified, followed by probability-weighted interpolation within a symmetric window of radius r centered at k_{max}:

$$D_t^l(u, v) = \frac{\sum_{k=k_{max}-r}^{k_{max}+r} D_{hypoth}(k) \cdot P_t(k, u, v)}{\sum_{m=k_{max}-r}^{k_{max}+r} P_t(m, u, v)}, \tag{17}$$

where $D_{hypoth}(k)$ denotes the pre-defined depth value at bin k. The resulting low-resolution depth map D_t^l is then upsampled to full resolution via a convex upsampling layer [23], yielding the final depth estimation D_t^m.

The proposed photometric gated cost volume (PGCV) systematically addresses the limitations of conventional cost volumes through its novel gating mechanism. Unlike traditional approaches, PGCV actively modulates feature representation at the preprocessing stage, suppressing low-variation regions while preserving discriminative features in high-texture areas and dynamic objects. our approach specifically targets low-texture areas (e.g., ground, sky) as demonstrated in Fig. 4. The resulting cost volume exhibits enhanced discriminability without additional computational burden, enabling more precise depth regression and improved robustness in complex dynamic environments.

E. Spectral Entropy Uncertainty Module

To address dynamic object issues and enhance depth regression using cost volume information, we propose the spectral entropy uncertainty module (SEUM) as shown in Fig. 3 (f). This module integrates spectral entropy with Fourier transform for uncertainty estimation and depth fusion. The Fourier transform converts spatial information into the frequency domain, facilitating the identification of noise introduced by dynamic objects. And spectral entropy analysis quantifies the complexity of these frequency components, allowing for a more precise characterization and management of uncertainty regions. We begin by computing the Fourier transform of the cost probability tensor P:

$$\hat{P} = \text{FFT}(P), \tag{18}$$

where $\text{FFT}(\cdot)$ denotes the Fast Fourier Transform operation. Next, we derive the magnitude spectrum S and normalize it to obtain a spectral probability distribution Q:

$$S = |\hat{P}|, \quad Q = \frac{S}{\sum S}. \tag{19}$$

We then calculate the spectral entropy H from Q:

$$H = -\sum Q \cdot \log(Q + \epsilon), \tag{20}$$

where $\epsilon = 10^{-9}$ is a small smoothing constant added to prevent numerical instability when Q approaches zero. A neural network is then employed to predict the uncertainty U from the computed spectral entropy H. Finally, we fuse the depth estimates D_{multi} and D_{mono} using the predicted uncertainty

TABLE I

DEPTH ESTIMATION RESULTS ON CITYSCAPES AND KITTI DATASETS. HERE, "TEST FRAMES" INDICATES THE NUMBER OF FRAMES USED FOR INFERENCE, WHILE "✓" DENOTES WHETHER A SEMANTIC SEGMENTATION MODEL IS UTILIZED.

	Method	Publication	Test Frames	Semantic	W×H	Errors ↓				Accuracy ↑		
						AbsRel	SqRel	RMSE	$\text{RMSE}_{\log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Cityscapes	Struct2Depth [24]	CVPR 2019	1	✓	416×128	0.145	1.737	7.280	0.205	0.813	0.942	0.976
	Monodepth2 [9]	ICCV 2019	1		416×128	0.129	1.569	6.876	0.187	0.849	0.957	0.983
	Gordon et al. [25]	ICCV 2019	1	✓	416×128	0.127	1.330	6.960	0.195	0.830	0.947	0.981
	Lee et al. [26]	ICCV 2021	1		832×256	0.116	1.213	6.695	0.186	0.852	0.951	0.982
	PPEA-Depth [27]	AAAI 2024	1		416×128	0.099	1.115	5.995	0.155	0.905	**0.976**	**0.991**
	Ours	-	1		416×128	**0.096**	**1.016**	**5.680**	**0.150**	**0.909**	**0.976**	**0.991**
	ManyDepth [6]	CVPR 2021	2(-1,0)		416×128	0.114	1.193	6.223	0.170	0.875	0.967	0.989
	DynamicDepth [7]	ECCV 2022	2(-1,0)	✓	416×128	0.103	1.000	5.867	0.175	0.895	0.967	0.991
	DS-Depth [11]	TCSVT 2023	2(-1,0)		416×128	0.100	1.055	5.884	0.155	0.899	0.974	0.991
	IterDepth [28]	TCSVT 2024	2(-1,0)		416×128	0.104	1.032	5.724	0.157	0.896	0.976	0.991
	Manydepth2 [29]	RAL 2025	2(-1,0)		416×128	0.097	**0.792**	5.827	0.154	0.903	0.975	**0.993**
	Ours	-	2(-1,0)		416×128	**0.093**	0.951	**5.559**	**0.147**	**0.910**	**0.977**	0.992
KITTI	Struct2Depth [24]	CVPR 2019	1	✓	640×192	0.141	1.026	5.291	0.215	0.816	0.945	0.979
	Monodepth2 [9]	ICCV 2019	1		640×192	0.115	0.903	4.863	0.193	0.877	0.959	0.981
	Lite-Mono [30]	CVPR 2023	1		640×192	0.101	0.729	4.454	0.178	0.897	0.965	0.983
	Wang et al. [31]	ICRA 2024	1		640×192	0.105	0.727	4.491	0.180	0.888	0.964	0.984
	GeoDepth [32]	CVPR 2025	1		640×192	0.100	0.694	4.381	0.176	0.897	0.966	0.984
	TCNet [33]	IROS 2025	1		640×192	**0.098**	0.726	4.295	0.174	**0.904**	**0.968**	0.984
	Ours	-	1		640×192	**0.098**	**0.675**	**4.284**	**0.173**	0.902	**0.968**	**0.985**
	ManyDepth [6]	CVPR 2021	2(-1,0)		640×192	0.098	0.770	4.459	0.176	0.900	0.965	0.983
	DynamicDepth [7]	ECCV 2022	2(-1,0)	✓	640×192	0.096	0.720	4.458	0.175	0.897	0.964	0.984
	MOVEDepth [34]	AAAI 2023	2(-1,0)		640×192	0.089	0.663	4.216	0.169	0.904	0.966	0.984
	MAL [10]	ICRA 2024	2(-1,0)		640×192	0.087	0.690	4.227	0.169	0.915	0.968	0.983
	Moon et al. [17]	CVPR 2024	2(-1,0)		640×192	0.096	0.696	4.327	0.174	0.904	0.968	**0.985**
	Manydepth2 [29]	RAL 2025	2(-1,0)		640×192	0.091	0.649	4.232	0.170	0.909	0.968	0.984
	TCNet [33]	IROS 2025	2(-1,0)		640×192	0.094	0.655	4.198	**0.164**	0.910	**0.970**	**0.985**
	Ours	-	2(-1,0)		640×192	**0.085**	**0.608**	**4.112**	**0.164**	**0.917**	0.969	**0.985**

U to obtain the fused depth D_{Fuse} following MOVEDepth [34]:

$$D_{Fuse} = (1 - U) \cdot D_{multi} + U \cdot D_{mono}. \tag{21}$$

The proposed strategy integrates single-frame and multi-frame depth estimates, dynamically adjusting the fusion process based on the spectral entropy uncertainty module. Additionally, the fused depth D_{Fuse} is used solely for loss computation, thereby not introducing additional inference overhead.

F. Training Objective

Our ReloDepth is trained in a self-supervised manner, and the loss consists of three parts:

$$\mathcal{L}_{total} = \mathcal{L}_{Mono}(D_{Mono}) + \mathcal{L}_{Multi}(D_{Multi}) + \mathcal{L}_{Fuse}(D_{Fuse}), \tag{22}$$

where $\mathcal{L}_{Mono}$, $\mathcal{L}_{Multi}$, and $\mathcal{L}_{Fuse}$ represent the loss terms associated with the depth estimates D_{Mono}, D_{Multi}, and D_{Fuse}, respectively. The loss function $\mathcal{L}(\cdot)$ is formulated as a weighted combination of the reprojection loss $\mathcal{L}_{rep}$ (Eq. 4) and the edge-aware smoothness loss $\mathcal{L}_s$ (Eq. 5) [20]:

$$\mathcal{L}(D) = \mathcal{L}_{rep}(D) \cdot \mathcal{M}_{relo} + \gamma \mathcal{L}_s(D), \tag{23}$$

where $\gamma = 0.001$ denotes the loss weight, and $\mathcal{M}_{relo}$ represents the proposed reprojection loss mask.

IV. EXPERIMENTS

We evaluate our method on two standard depth estimation benchmarks: KITTI [35] and Cityscapes [36]. Following the Eigen split, we use 39,810 images for training, 4,424 for validation, and 697 for testing. Cityscapes contains more dynamic objects, providing a challenging testbed for dynamic scenes. We use 69,731 monocular triplets for training and 1,525 for testing, capping depth at 80 meters. Models are implemented in PyTorch and trained on NVIDIA RTX 4090 GPUs. Training uses batch size 12 for 20 epochs on KITTI and 5 epochs on Cityscapes.

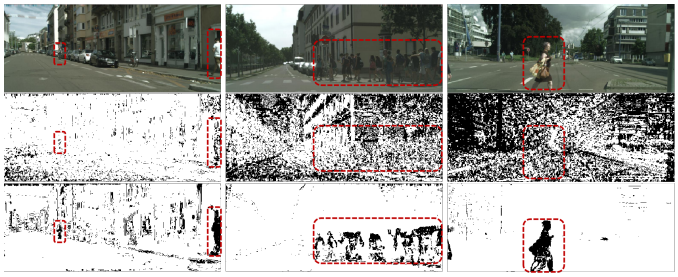

Fig. 5. Visualization of the proposed reprojection loss mask. As indicated by the red boxes, the reprojection loss mask (the third row) successfully masks dynamic objects such as pedestrians, which are not properly handled by the conventional auto-masking method [9] (the second row) that only removes pixels static relative to the camera motion.

TABLE II

ABLATION STUDY ON CITYSCAPES.

ReloMask		PGCV	SEUM	Errors ↓		Accuracy ↑	
Single	Multi			AbsRel	RMSE	$\delta < 1.25$	$\delta < 1.25^2$
				0.102	5.792	0.892	0.973
✓				0.099	5.793	0.903	0.975
✓	✓			0.099	5.768	0.907	0.975
		✓		0.099	5.777	0.895	0.976
			✓	0.101	5.721	0.896	0.974
✓	✓	✓		0.096	5.594	0.907	**0.977**
✓	✓	✓	✓	**0.093**	**5.559**	**0.910**	**0.977**

A. Comparison with State-of-the-Art Methods

Our method is evaluated on the KITTI and Cityscapes benchmarks, with quantitative results presented in Table I. On Cityscapes, characterized by highly dynamic and complex urban environments, our approach demonstrates substantial im-

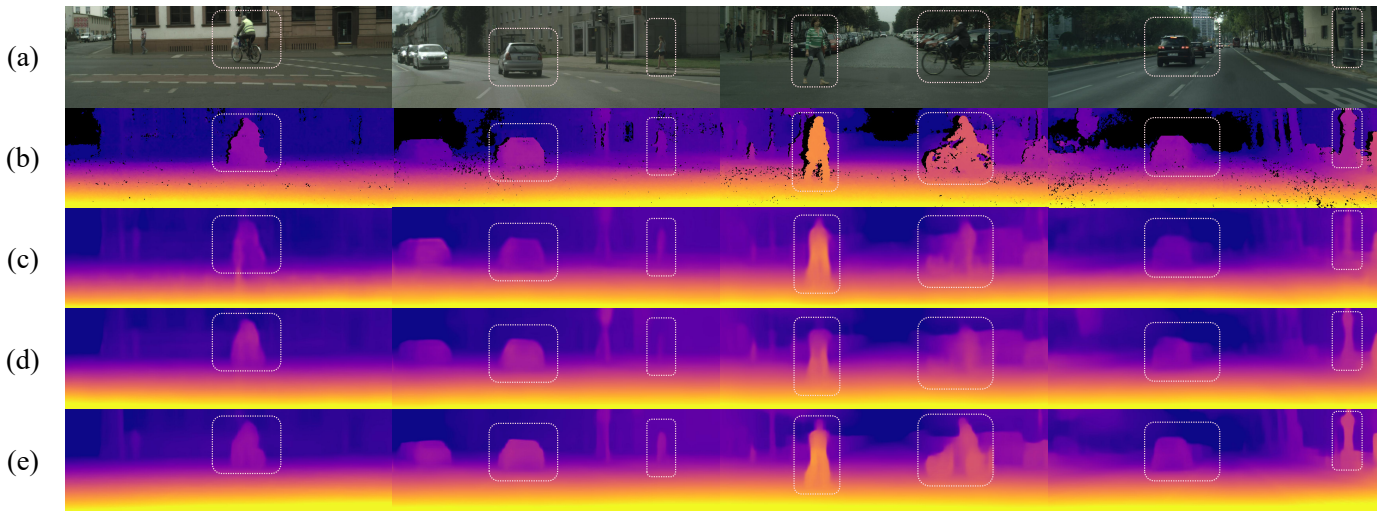

Fig. 6. Qualitative results of ReloDepth on the Cityscapes [36] dataset, where (a) is the input image, (b) is the ground truth, and (c), (d), and (e) represent the outputs of Monodepth2 [9], ManyDepth [6], and our method, respectively. Our approach achieves better accuracy, particularly in dynamic regions, as highlighted in the pink boxes.

TABLE III
ABLATION STUDY ON CITYSCAPES **DYNAMIC REGIONS**. EVALUATION FOLLOWS THE MASKED-REGION PROTOCOL OF [7] TO ISOLATE MOVING OBJECTS.

ReloMask	PGCV	SEUM	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	#Params	MACs
			0.126	1.388	4.820	0.175	23.86M	22.265G
✓			0.119	1.420	4.725	0.167	23.86M	22.265G
✓	✓		0.116	1.235	4.460	0.163	23.86M	22.266G
✓	✓	✓	**0.114**	**1.218**	**4.435**	**0.162**	23.86M	22.266G

TABLE IV
ABLATION ON THE INTEGRATION OF RELOMASK WITH MONODEPTH2 AND MANYDEPTH.

Method	Train	Dataset	Errors ↓		Accuracy ↑	
			AbsRel	RMSE	$\delta < 1.25$	$\delta < 1.25^2$
Monodepth2 [9]	M	Cityscapes	0.129	6.876	0.849	0.957
Ours-Monodepth2 with ReloMask	M	Cityscapes	**0.117**	**6.432**	**0.875**	**0.966**
ManyDepth [6]	M	Cityscapes	0.114	6.223	0.875	0.967
Ours-ManyDepth with ReloMask	M	Cityscapes	**0.110**	**6.129**	**0.892**	**0.970**
Monodepth2 [9]	S	KITTI	**0.109**	4.960	0.864	0.948
Ours-Monodepth2 with ReloMask	S	KITTI	**0.109**	**4.959**	**0.874**	**0.954**

provements over state-of-the-art methods, achieving a $\delta < 1.25$ accuracy of 91%. On the KITTI dataset, where dynamic objects account for only 0.34% of pixels [7], [10] and most vehicles are stationary, our method still attains state-of-the-art performance, with a $\delta < 1.25$ of 0.917, significantly outperforming existing methods designed for dynamic scenes. These results confirm the robustness and generalizability of our approach across diverse driving conditions. Notably, our approach exhibits compelling performance even on individual frames, exceeding the capabilities of current state-of-the-art single-frame methods and performing favorably when benchmarked against numerous multi-frame strategies. Qualitative results in Fig. 6 show that our method produces depth maps with sharper object boundaries and more consistent predictions for dynamic objects.

B. Ablation Study

We evaluate the contributions of each component in ReloDepth on the Cityscapes dataset, with detailed performance and dynamic region analyses provided in Table II and Table III, respectively. In Table II, the proposed ReloMask consistently enhances both single-frame and multi-frame depth estimation, achieving a competitive $\delta < 1.25$ of 0.907 in the combined setting. Table IV validates the architecture-agnostic effectiveness of ReloMask. Its integration into Monodepth2 and ManyDepth yields notable improvements. Specifically, under the Monodepth2 stereo training protocol, ReloMask elevates $\delta < 1.25$ from 0.864 to 0.874. In dynamic regions, ReloMask reduces Abs Rel from 0.126 to 0.119, further optimized to 0.114 with PGCV and SEUM (Table III). Visual comparisons in Fig. 5 confirm its ability to accurately isolate moving objects like pedestrians without semantic supervision. Remarkably, our full model maintains high efficiency with 23.86M parameters and 22.266G MACs (with only a 0.001G increment from PGCV), achieving SOTA results with minimal overhead.

V. CONCLUSION

In this paper, we propose ReloDepth, a novel self-supervised depth estimation framework specifically designed to tackle challenges posed by dynamic objects. We introduce the reprojection loss mask to mitigate the impact of dynamic objects on the loss function, improving robustness in handling dynamic entities. Additionally, we refine multi-frame depth estimation through the photometric gated cost volume strategy, which proactively eliminates redundant regions within the cost volume, thereby optimizing the focus and enhancing the robustness of the depth estimation process. Furthermore, we incorporate the spectral entropy uncertainty module to guide depth fusion and address challenges associated with dynamic objects in the cost volume. Our method achieves state-of-the-art results on the KITTI and Cityscapes datasets.

REFERENCES

[1] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8445–8453.

[2] J. Biswas and M. Veloso, "Depth camera based indoor mobile robot localization and navigation," in *IEEE International Conference on Robotics and Automation*, 2012, pp. 1697–1702.

[3] F. Flacco, T. Kröger, A. De Luca, and O. Khatib, "A depth space approach to human-robot collision avoidance," in *IEEE International Conference on Robotics and Automation*, 2012, pp. 338–345.

[4] L. Sun, J.-W. Bian, H. Zhan, W. Yin, I. Reid, and C. Shen, "Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, pp. 497–508, 2023.

[5] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1851–1858.

[6] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1164–1174.

[7] 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 *Proceedings of the European Conference on Computer Vision*, 2022, pp. 228–244.

[8] J. Liu, L. Kong, B. Li, Z. Wang, H. Gu, and J. Chen, "Mono-vifi: A unified learning framework for self-supervised single and multi-frame monocular depth estimation," in *Proceedings of the European Conference on Computer Vision*, 2024, pp. 90–107.

[9] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.

[10] Y.-J. Dong, F.-L. Zhang, and S.-H. Zhang, "Mal: Motion-aware loss with temporal and distillation hints for self-supervised depth estimation," in *IEEE International Conference on Robotics and Automation*, 2024, pp. 7318–7324.

[11] X. Miao, Y. Bai, H. Duan, Y. Huang, F. Wan, X. Xu, Y. Long, and Y. Zheng, "Ds-depth: Dynamic and static depth estimation via a fusion cost volume," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[12] R. Marsal, F. Chabot, A. Loesch, W. Grolleau, and H. Sahbi, "Monoprob: Self-supervised monocular depth estimation with interpretable uncertainty," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 3637–3646.

[13] Y. Wang, Y. Liang, H. Xu, S. Jiao, and H. Yu, "Sqldepth: Generalizable self-supervised fine-structured monocular depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 5713–5721.

[14] J. Wang, C. Lin, L. Nie, S. Huang, Y. Zhao, X. Pan, and R. Ai, "Weatherdepth: Curriculum contrastive learning for self-supervised depth estimation under adverse weather conditions," in *IEEE International Conference on Robotics and Automation*, 2024, pp. 4976–4982.

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

[16] V. Guizilini, R. Ambruș, D. Chen, S. Zakharov, and A. Gaidon, "Multi-frame self-supervised depth with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 160–170.

[17] J. Moon, J. L. G. Bello, B. Kwon, and M. Kim, "From-ground-to-objects: Coarse-to-fine self-supervised monocular depth estimation of dynamic objects with ground contact prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 519–10 529.

[18] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[20] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[21] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3273–3282.

[22] F. Wang, S. Galliani, C. Vogel, and M. Pollefeys, "Itermvs: Iterative probability estimation for efficient multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8606–8615.

[23] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 402–419.

[24] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Unsupervised monocular depth and ego-motion learning with structure and semantics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1–8.

[25] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8977–8986.

[26] S. Lee, F. Rameau, F. Pan, and I. S. Kweon, "Attentive and contrastive learning for joint depth and motion field estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4862–4871.

[27] Y.-J. Dong, Y.-C. Guo, Y.-T. Liu, F.-L. Zhang, and S.-H. Zhang, "Ppea-depth: Progressive parameter-efficient adaptation for self-supervised monocular depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 1609–1617.

[28] C. Feng, Z. Chen, C. Zhang, W. Hu, B. Li, and F. Lu, "Iterdepth: Iterative residual refinement for outdoor self-supervised multi-frame monocular depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 329–341, 2023.

[29] K. Zhou, J.-W. Bian, J.-Q. Zheng, J. Zhong, Q. Xie, N. Trigoni, and A. Markham, "Manydepth2: Motion-aware self-supervised monocular depth estimation in dynamic scenes," *IEEE Robotics and Automation Letters*, 2025.

[30] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 537–18 546.

[31] C. Wang, G. Zhang, and W. Zhou, "Self-supervised learning of monocular visual odometry and depth with uncertainty-aware scale consistency," in *2024 IEEE International Conference on Robotics and Automation*. IEEE, 2024, pp. 3984–3990.

[32] H. Wu, S. Gu, L. Duan, and W. Li, "Geodepth: From point-to-depth to plane-to-depth modeling for self-supervised monocular depth estimation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11 525–11 535.

[33] Y. Zhu, H. Liu, J. Wu, and M. Liu, "Tcnet: A temporally consistent network for self-supervised monocular depth estimation," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2025, pp. 678–685.

[34] X. Wang, Z. Zhu, G. Huang, X. Chi, Y. Ye, Z. Chen, and X. Wang, "Crafting monocular cues and velocity guidance for self-supervised multi-frame depth learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 2689–2697.

[35] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[36] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.