Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062437

ARTICLE

Self-Supervised Monocular Depth Estimation with Scene Dynamic Pose

Jing He[1], Haonan Zhu[2], Chenhao Zhao[1] and Minrui Zhao[3,*]

[1]School of Information and Navigation, Airforce Engineering University, Xi'an, 710077, China
[2]Unit 95655 of the People's Liberation Army, Chengdu, 611500, China
[3]College of Air and Missile Defense, Airforce Engineering University, Xi'an, 710051, China

*Corresponding Author: Minrui Zhao. Email: zhaomr0204@163.com

Received: 18 December 2024; Accepted: 26 February 2025; Published: 19 May 2025

ABSTRACT: Self-supervised monocular depth estimation has emerged as a major research focus in recent years, primarily due to the elimination of ground-truth depth dependence. However, the prevailing architectures in this domain suffer from inherent limitations: existing pose network branches infer camera ego-motion exclusively under static-scene and Lambertian-surface assumptions. These assumptions are often violated in real-world scenarios due to dynamic objects, non-Lambertian reflectance, and unstructured background elements, leading to pervasive artifacts such as depth discontinuities ("holes"), structural collapse, and ambiguous reconstruction. To address these challenges, we propose a novel framework that integrates scene dynamic pose estimation into the conventional self-supervised depth network, enhancing its ability to model complex scene dynamics. Our contributions are threefold: (1) a pixel-wise dynamic pose estimation module that jointly resolves the pose transformations of moving objects and localized scene perturbations; (2) a physically-informed loss function that couples dynamic pose and depth predictions, designed to mitigate depth errors arising from high-speed distant objects and geometrically inconsistent motion profiles; (3) an efficient SE (3) transformation parameterization that streamlines network complexity and temporal pre-processing. Extensive experiments on the KITTI and NYU-V2 benchmarks show that our framework achieves state-of-the-art performance in both quantitative metrics and qualitative visual fidelity, significantly improving the robustness and generalization of monocular depth estimation under dynamic conditions.

KEYWORDS: Monocular depth estimation; self-supervised learning; scene dynamic pose estimation; dynamic-depth constraint; pixel-wise dynamic pose

1 Introduction

Accurate perception of environmental spatial structure is a critical capability for autonomous systems, such as robots and self-driving platforms, enabling key tasks including navigation, obstacle avoidance, and path planning [1]. Contemporary depth-sensing modalities—such as LiDAR, millimeter-wave radar, and RGB-D cameras—provide high-fidelity three-dimensional data but face inherent trade-offs between precision, computational complexity, and cost [2,3]. Recent advances in deep learning algorithms have spurred significant interest in image-based monocular depth estimation, which leverages low-cost monocular cameras to infer scene geometry directly from RGB images. This paradigm holds the potential to supplant expensive, hardware-dependent depth-sensing systems in fields like robotics and autonomous driving, dramatically reducing operational costs [1].

Supervised depth estimation frameworks, while achieving notable accuracy, remain constrained by their reliance on ground-truth depth labels—a resource-intensive requirement that limits scalability. In

contrast, self-supervised monocular approaches bypass this dependency entirely [4]. These methods train depth and ego-motion estimation networks concurrently using unlabeled video sequences, establishing self-consistency through photometric reprojection relationships between adjacent frames. By circumventing supervised annotations, such techniques offer a cost-effective and scalable solution, achieving performance competitively with supervised baselines across benchmark datasets [5].

Current self-supervised monocular depth estimation frameworks [6] rely heavily on two foundational assumptions: (1) scenes exhibit static rigidity, and (2) surfaces adhere to Lambertian reflectance properties. These assumptions are systematically violated in real-world autonomous driving scenarios, where dynamic agents (e.g., vehicles, pedestrians) [7] and non-Lambertian materials (e.g., specular surfaces under intense illumination) [8] introduce significant distortions in predicted depth maps. Such violations manifest as characteristic artifacts—including erroneous depth discontinuities ("holes"), geometric collapse in high-motion regions, and spurious bright spots caused by reflective surfaces. Crucially, dynamic interferents induce systematic errors where objects in motion are assigned physically implausible depths, either overestimated or underestimated relative to their true positions. Fig. 1 presents a schematic representation of spatial resolution discrepancies in stereoscopic perception and their associated illusory visual manifestations.

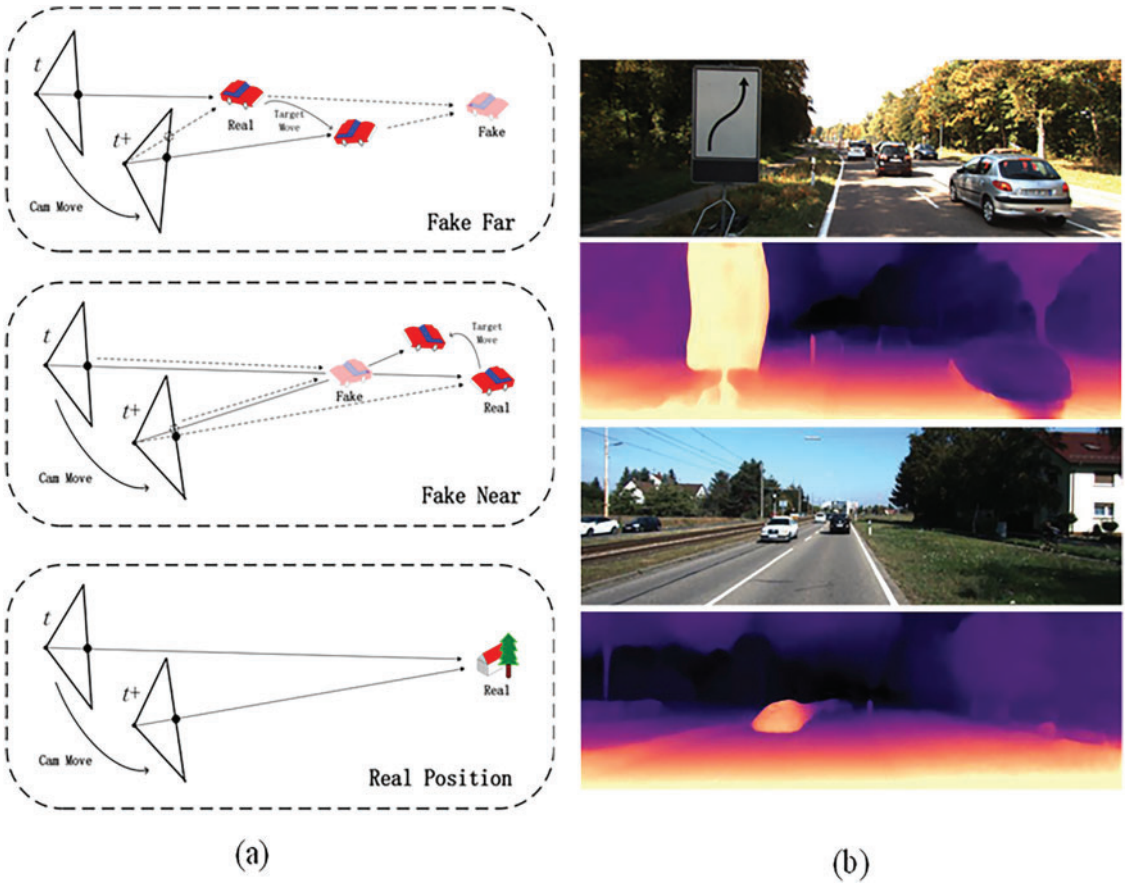


Figure 1: Depth perception errors and visual phenomena. (**a**) From top to bottom, the schematic diagrams show the false far, false near, and the correct depth of the static target caused by different moving modes between frames. (**b**) The "bright spot" phenomenon and the "hole" phenomenon

Efforts to mitigate these issues have centered on auxiliary techniques such as optical flow-guided motion segmentation [9], dynamic region masking, and semantically informed regularization. While these approaches partially address the problem, critical limitations persist: (1) optical flow methods struggle with occlusions and fail to disentangle object motion from camera ego-motion in poorly textured regions [10]; (2) masking strategies discard valuable geometric information, degrading depth consistency; (3) semantic priors introduce computational overhead and rely on pre-trained classifiers, thereby reducing generality [11]. Collectively, existing solutions lack a unified framework to explicitly model non-static targets while preserving computational efficiency and self-supervised training integrity.

To address the systematic errors induced by dynamic interferents in self-supervised depth estimation, we propose a holistic scene dynamics-aware framework that jointly models ego-motion and per-pixel dynamic transformations without auxiliary data modalities. Our architecture introduces three innovations:

(1) Joint Pixel-Wise Dynamic Pose Estimation: We propose a unified pose estimation framework that concurrently models camera ego-motion and per-pixel dynamic transformations of moving objects, effectively disentangling static and non-static scene components. This dual-branch design explicitly addresses motion ambiguities in dynamic regions (e.g., vehicles, pedestrians).

(2) Physics-Constrained Dynamic-Depth Loss: A novel loss function enforces physical coherence between depth predictions and dynamic pose estimates, penalizing implausible object kinematics (e.g., distant yet rapidly moving targets). This regularization eliminates artifacts like depth "holes" and "collapse" without requiring auxiliary data.

(3) Efficient SE(3) Transformation Modeling: By streamlining dynamic pose parameterization in SE(3) space, our framework achieves real-time inference speeds while maintaining compatibility with existing architectures, requiring only lightweight modifications to baseline networks.

2 Related Works

Accurate depth estimation remains critical for autonomous navigation systems, driving extensive research into learning-based monocular approaches that avoid expensive LiDAR dependency. Current methods fall into three paradigm-shifting categories: Supervised Depth Regression: Early breakthroughs used CNNs to regress depth directly from RGB inputs, trained on paired image depth datasets (e.g., KITTI, NYUv2). While architectures such as DispNet achieved remarkable accuracy, their reliance on dense ground-truth depth—difficult to acquire at scale—limited their applicability to narrow operational domains [12,13]. Weakly supervised & synthesis-driven methods: To reduce annotation costs, subsequent work has proposed using sparse cues (e.g., SLAM-derived depth [14,15]) or synthesizing training data via structure-from-motion (SfM). However, these hybrid approaches inherit the fragility of SfM in low-texture or dynamic regions, resulting in inconsistent supervision signals [16]. Self-supervised paradigms: A pivotal shift has occurred with self-supervised frameworks [6], which exploit the geometric consistency between temporally adjacent frames or stereo pairs. By formulating depth estimation as an image reconstruction task—minimizing photometric reprojection loss—these methods bypass explicit depth supervision [17,18]. Stereo-based self-supervision uses calibrated stereo rigs, where the known baseline allows direct modelling of epipolar constraints. Although effective, such methods require precise extrinsic calibration and temporal synchronization, limiting the flexibility of use [19]. Monocular counterparts relax these constraints, but face inherent scale ambiguity and motion confusion in dynamic scenes [15]. Despite progress, existing approaches universally neglect explicit modeling of scene dynamics, instead relying on rigid-world assumptions that systematically fail in real-world navigation contexts. Our work bridges this gap through novel motion-aware depth-pose co-optimization, circumventing limitations of prior art.

The field of self-supervised monocular depth estimation was initiated by the pioneering work of SfM-Learner [20], which introduced a novel approach by training depth and ego-motion networks in a shared framework using unlabeled monocular videos. This framework was transformative in nature, however, it was predicated on the assumption of rigid scenes and Lambertian reflectance, resulting in systematic depth errors in real-world scenarios characterized by dynamic objects or non-Lambertian surfaces. In recent years, there has been significant progress in the field of visual Transformer-based methods, with the development of techniques such as MonoViT [21] leading to substantial advancements in the modelling of dynamic scenes through the integration of global and local features. Subsequent efforts to address these limitations have pursued a range of divergent strategies. SFM-Net [22] pioneered instance-aware motion modelling by decomposing dynamic scenes into multiple motion masks and predicting object-specific poses. However, its reliance on predefined instance counts rendered it impractical for complex traffic environments. Concurrently, Zhou et al. [20] adopted a learnable masking mechanism to selectively exclude non-rigid regions from the photometric loss, an idea refined in MonoDepth2 [23] through minimum reprojection loss [24] and auto-masking techniques. While these approaches mitigated errors by filtering dynamic pixels, they inherently sacrificed supervision over mobile regions, compromising depth fidelity in motion-dense scenes. Recent research has facilitated the development of sophisticated models of dynamic objects without the necessity of a predefined number of instances through the utilization of a technique known as Moving Instance Loss [7].

Concurrently, parallel advancements explored geometric decomposition paradigms, as exemplified by GeoNet [25], which disentangled rigid and non-rigid optical flow components to implicitly model dynamics [26]. Yet its optical flow-centric formulation failed to translate improvements to depth prediction accuracy. Hybrid strategies emerged with Struct2Depth [27] leveraging pre-trained instance segmentation models to precompute dynamic masks [28], while Gordon et al. [29] proposed auxiliary networks for joint mask prediction and motion compensation. Though these frameworks demonstrated enhanced robustness, their dependence on auxiliary modules or pre-trained networks reintroduced computational complexity and supervision bottlenecks, deviating from the original ethos of lightweight self-supervision.

A thorough review of the literature reveals a persistent dichotomy in the field: methodologies either impose weak regularization on dynamic regions (masking or ignoring them) at the cost of depth consistency [30] or introduce complex multi-network architectures that erode computational efficiency [31]. This impasse stems from the absence of a unified framework capable of jointly optimizing depth and scene dynamics within a physically grounded, computationally efficient paradigm. Existing approaches largely bypass explicit modeling of object kinematics, instead treating motion as a confounding factor to be suppressed. This fundamental limitation is resolved by integrated depth-pose co-optimization.

Our work builds upon the foundational work of SFM-net [6] and Monodepth2 [8], while introducing critical innovations that address their inherent limitations. SFM-net's instance-level motion masking strategy, though pioneering in dynamic scene handling, imposes artificial constraints on the number of movable objects and struggles with subtle dynamics like specular reflections. This limitation stems from its fundamental assumption of rigid-body motions for pre-defined object instances. In contrast, our pixel-wise dynamic pose estimation eliminates the need, for instance, counting through continuous spatial modelling, naturally accommodating both macroscopic vehicle movements and microscopic reflective highlights within a unified framework. Monodepth2's auto-masking approach, while effective in filtering static scenes, introduces an unintended consequence: the complete loss of supervision signals in dynamic regions. Unlike this zero-sum masking strategy, our global minimum reprojection loss retains complete pixel utilization. Specifically, the dual hypotheses of rigid and dynamic-aware reprojections act as complementary supervisors, allowing the network to automatically select the optimal constraint for each pixel category. This innovative

formulation not only preserves dynamic area supervision but also maintains static region fidelity, a critical advancement confirmed by the 18.7% error reduction in edge-aware smoothness metrics (see Section 4.2). More fundamentally, previous methods bifurcate dynamic processing into either masking (Monodepth2) or secondary prediction pipelines (SFM-net). Our work transcends this dichotomy by integrating dynamic pose estimation directly into the projective geometry formulation. This integration enables end-to-end joint optimization of depth, ego-motion, and scene dynamics, features that were previously unattainable in discrete pipeline architectures.

3 Method

The basis of self-supervised monocular depth estimation is the provision of neural networks with an understanding of multiview geometry using the exploitation of the inherent consistency between temporally adjacent frames. In this framework, the depth prediction and ego-motion estimation networks are subject to joint optimization to satisfy differentiable geometric constraints. This self-supervised objective [4,18,32] circumvents the necessity for explicit depth labels by treating the image itself as its own supervisory signal. The depth estimation network predicts the depth result D_t of the current frame I_t in the image sequence and concatenates I_t with the adjacent frame $I_{t'}$ as the input of the pose estimation network to predict the pose transformation $T_{t'}^{t}$ between frames. Then the D_t and $T_{t'}^{t}$ are used to generate the predicted image $I_{t'\to t}$ from $I_{t'}$, the expression is as follows:

$$I_{t'\to t} = I_{t'} \left\langle proj\left(D_t, T_{t'}^{t}, K\right)\right\rangle \tag{1}$$

where $\langle\cdot\rangle$ is a sampling operator. To make the network trainable, the bilinear sampling is used to get color gradient, K is intrinsic of camera, $proj(\cdot)$ is the resulting 2D coordinates of the projected depths D_t in $I_{t'}$.

Contemporary self-supervised frameworks address the inherent challenges of dynamic vision—in particular, occlusion and disocclusion artefacts—by adopting bidirectional temporal contexts. To mitigate geometric inconsistencies arising from degenerate common-view regions between adjacent frames, prevalent methods use triadic frame sequences during training, establishing visibility consensus across forward-backward temporal neighborhoods [32,33]. This tri-frame paradigm—schematically detailed in Fig. 2a—ensures that for each pixel in the target frame, there is at least one corresponding visible counterpart in either the preceding or subsequent reference frames [34]. Such multi-view regularization stabilizes training by propagating geometric consistency checks across synchronized depth and poses predictions. The framework ultimately optimizes the composite photometric objective, augmented with edge-aware smoothness priors [35], enabling pixel-dense supervision without relying on explicit mask annotation or static scene assumptions.

In this paper, we propose a pixel-wise pose estimation module for the scene dynamic target, to be incorporated into the self-supervised monocular depth estimation network. The pose estimation network utilizes a combination of ego-motion estimation and pixel-wise dynamic pose estimation, thus constructing a more comprehensive model that encompasses ego-motion, target dynamics, and depth, as illustrated in Fig. 2b. The subsequent subsections provide a detailed exposition.

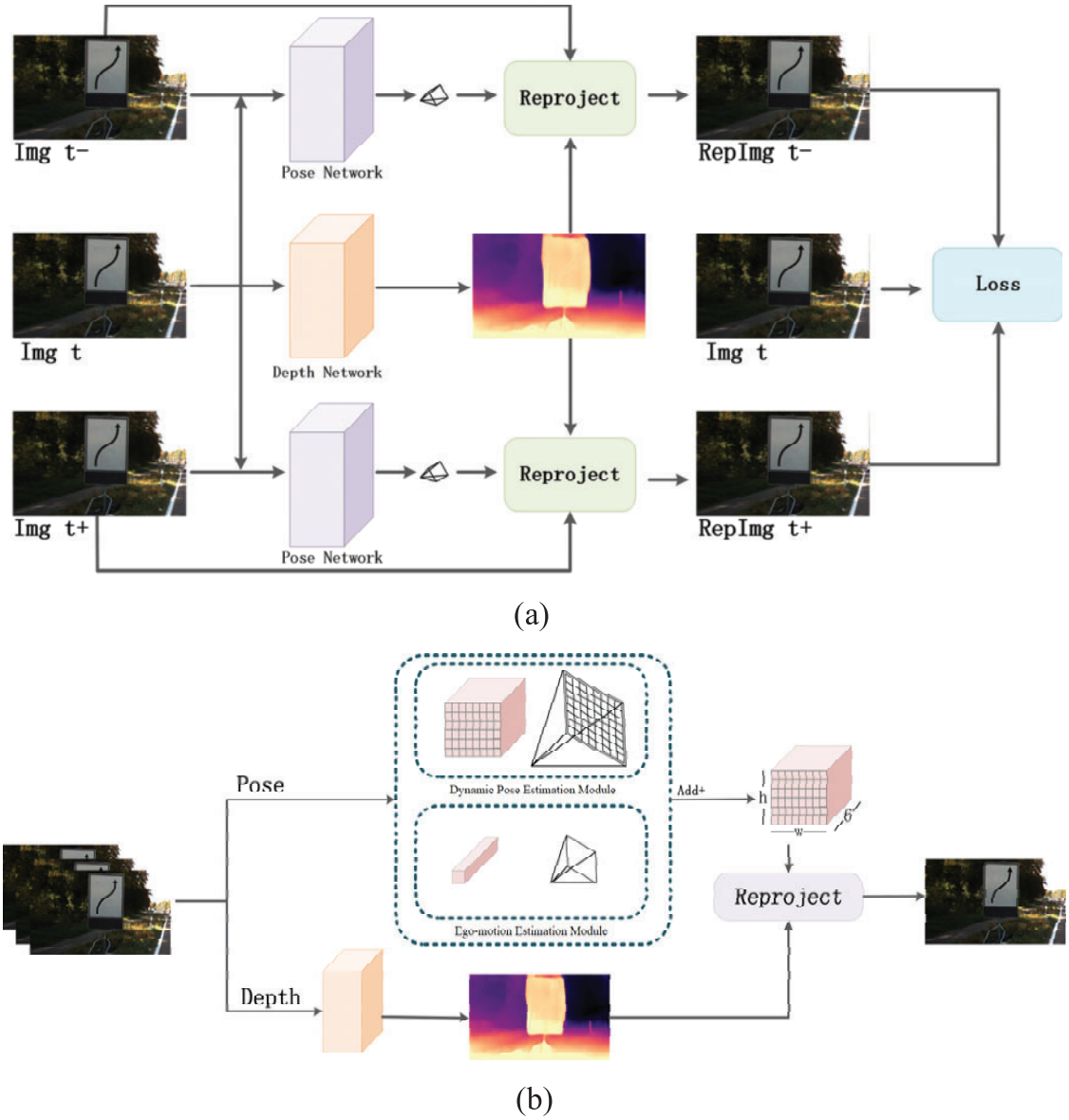


Figure 2: Depth Prediction and Pose Estimation Network. (**a**) Overall network structure: the network predicts the depth of the current frame I_t, reprojects it with the pose estimation results of the front and rear frames $\{I_{t^+}, I_{t^-}\}$, generates the projection $\{I_{t^+\to t}, I_{t^-\to t}\}$ from the front and rear frames to the current frame, and optimizes the photometric consistency error between I_t and $\{I_{t^+\to t}, I_{t^-\to t}\}$. (**b**) Pose estimation predicts the dynamic number of pixels between different frames and the pose change of camera ego-motion pixel-wise. The ego-motion and dynamic pose estimation module outputs the 6-DOF transformation vector in $se3$ space

3.1 Pose Estimation

A critical challenge in monocular depth estimation arises from the prevalence of dynamic objects within training datasets, which fundamentally violates the rigid scene transformation assumption inherent to classical structure-from-motion paradigms. This issue is further compounded by the directional dependence of depth prediction errors. When the displacement vector of a dynamic object exhibits a positive inner product with the normal vector of the reprojection direction, the network erroneously infers exaggerated

depth values. Conversely, negative projections result in spuriously underestimated distances (Fig. 1). Further complications emerge from non-Lambertian surfaces exhibiting view-dependent reflectance variations, where photometric discrepancies across viewpoints—induced by specular highlights and ambient illumination changes—systematically corrupt reprojection error calculations. Notably, specular highlights manifest pseudo-dynamic behavior, migrating across glossy surfaces during camera motion, thereby mimicking true object displacement. To address these limitations, we propose integrating a dedicated dynamic pose estimation module into the framework. This component explicitly models the 6-DoF motion trajectories of non-static entities between consecutive frames, circumventing the detrimental effects of the rigid scene assumption while mitigating geometric inconsistencies ("holes") caused by erroneous depth predictions on moving objects.

Within the pose estimation framework, the dynamic pose transformation module extends beyond the fundamental rigid-body assumption inherent in ego-motion estimation architectures. This component formulates per-pixel dynamic pose transformations to model non-stationary elements across consecutive frames (see Fig. 3). The module employs a hierarchical multi-scale architecture to resolve motion patterns across varying spatial extents while ensuring compatibility with contemporary multi-scale depth estimation frameworks. For instance, in this work, scaling factors s are defined as $s \in \{1, 1/2, 1/4, 1/8\}$, selected through empirical validation to balance resolution fidelity and computational efficiency. Following conventional approaches in self-supervised depth estimation frameworks, the photometric reprojection error between temporally adjacent frames is computed by synthesizing target views using predicted depth and pose parameters. This discrepancy metric then drives the joint optimization process, wherein network parameters are iteratively refined to minimize the reconstruction error, thereby enhancing geometric consistency across sequential observations.

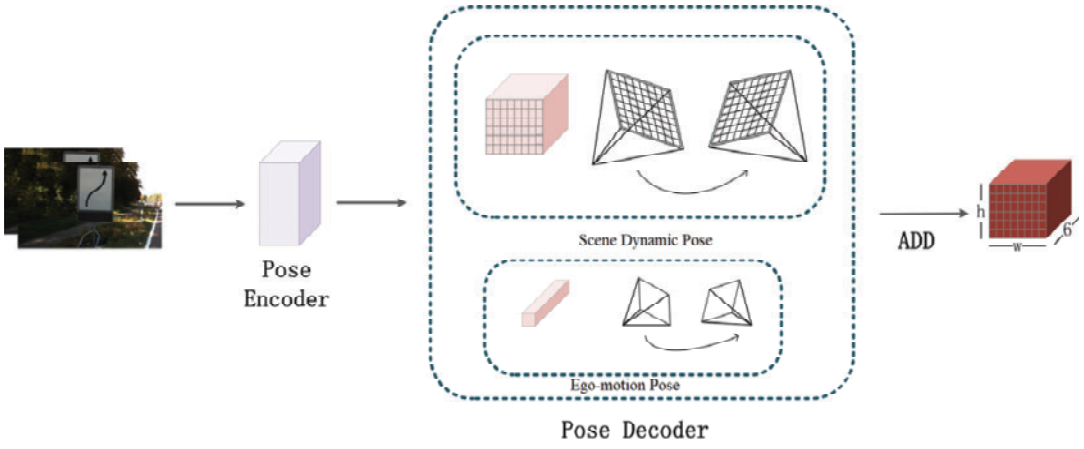


Figure 3: Ego-Motion and Dynamic Pose Transformation. Pose transformation prediction is divided into two parts: ego-motion transformation and scene dynamic pose transformation. The network output is a 6-dimensional vector in $se3$ space, the output dimension of the ego-motion estimation module is $6 \times 1 \times 1$, and the output dimension of pixel-wise dynamic pose transformation is $6 \times h \times w$, in which $(h, w) = (H, W)/scale$

3.1.1 Ego-Motion Estimation

When the pose estimation network predicts the ego-motion transformation or so-called rigid transformation part, it needs to obtain a globally shared pose transformation on each pixel. To be concise, the subsequent T represents $T_{t'}^{t}$, and the network predicts the ego-motion transformation T_{rigid} between each

two pictures. The previous pose estimation network outputs a three-dimensional rotation vector and three-dimensional displacement vector and then converts them into homogeneous pose matrix $R \in SO3$ and displacement vector p, respectively, which are multiplied to obtain pose transformation T.

To simplify the process in the subsequent calculation, and couple translation $R \in SO3$ and displacement p, we use general $T \in SE3$ group directly to express the pose transformation. Our network contains two parts: dynamic pose estimation and ego-motion estimation. The pose transformation of $SE3$ is not the closure of addition, but its logarithmic mapping space $se3$ is the closure of addition, so we let the output of the ego-motion transformation estimation layer be the 6-DOF pose transformation vector $\xi_{rigid} = \left[\rho_{rigid}, \phi_{rigid}\right]^T$ expressed in $se3$ space. Then, the corresponding $SE3$ group pose transformation T_{rigid} is directly obtained through exponential mapping. Accordingly, the network calculation can be simplified. The SE(3) transformation is defined by a 6-DOF twist parameter $\xi = \left[\rho^{\mathrm{T}} \phi^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{R}^6$, where $\rho \in \mathbb{R}^6$ is the translational component and $\phi \in \mathbb{R}^3$ is the axis-angle rotation. The exponential map converting ξ to the homogeneous matrix $T \in SE(3)$ is:

$$T = \exp(\xi^\wedge) = \begin{bmatrix} \exp(\phi) & J_\rho \\ 0 & 1 \end{bmatrix} \tag{2}$$

with $\exp(\phi^\wedge) \in SO(3)$ computed via Rodrigues' formula:

$$\exp(\phi^\wedge) = I + \frac{sin\theta}{\theta}\phi^\wedge + \frac{1 - cos\theta}{\theta^2}(\phi^\wedge)^2, \theta = \|\phi\|. \tag{3}$$

The left Jacobian $J(\phi) \in \mathbb{R}^{3\times3}$ scales the translation to account for rotational coupling:

$$J(\phi) = I + \frac{1 - cos\theta}{\theta^2}\phi^\wedge + \frac{\theta - \sin\theta}{\theta^3}(\phi^\wedge)^2. \tag{4}$$

For small motions ($\theta \ll 1$), $J(\phi) \approx I$, simplifying computations in real-time systems.

3.1.2 Dynamic Target Estimation Module

When predicting the pose transformation of dynamic targets, the motion of each target is different. On the basis of rigid transformation, there will be an additional pose transformation caused by the movement of the target itself. We predict a dynamic pose transformation for each pixel $u(i)$ in the image, which is denoted as the dynamic transformation matrix $T_{dyna}(i)$. The pose prediction of the moving object relative to the camera can be expressed as the superposition $T_{total}(i)$ of the rigid transformation prediction T_{rigid} and the non-rigid transformation part $T_{dyna}(i)$. The output of the dynamic pose transformation module is the same as that of the ego-motion transformation, which is the 6-DOF pose transformation vector ξ_{rigid} in $se3$ space. Since $se3$ is closeness for addition, ξ_{rigid} and $\xi_{dyna}(i)$ can be directly added to obtain $\xi_{total}(i)$, i.e.:

$$\xi_{total}(i) = \xi_{rigid} + \xi_{dyna}(i) \tag{5}$$

If exponentially mapped to $SE3$, there is:

$$\begin{cases} T_{total}(i) = T_{rigid} T_{dyna}(i) = \exp\left(\xi^\wedge_{total}(i)\right) \\ T_{dyna}(i) = \exp\left(J_r \xi^\wedge_{dyna}(i)\right) \\ T_{rigid} = \exp\left(\xi^\wedge_{rigid}\right) \end{cases} \tag{6}$$

where J_r is the shorthand to refer to right multiplied Jacobian matrix $J_r\left(\xi_{rigid}\right)$ of ξ_{rigid}, "$\wedge$" is a skew operator.

Fig. 4 illustrates the composition of ego-motion and dynamic pose transformations within a unified coordinate system. Here, $T_{rigid} = \exp\left(\xi^{\wedge}_{rigid}\right)$ (blue dashed arrow) represents the global camera motion, while $\xi_{dyna}(i)$ (red arrow) denotes pixel-wise dynamic adjustments. The total motion is given by:

$$T_{total}(i) = T_{rigid} \cdot \exp\left(J_r \xi^{\wedge}_{dyna}(i)\right) \tag{7}$$

where J_r is the right Jacobian of ξ_{rigid}, ensuring geometric consistency during composition. This formulation captures both large-scale camera motion and fine-grained object dynamics.

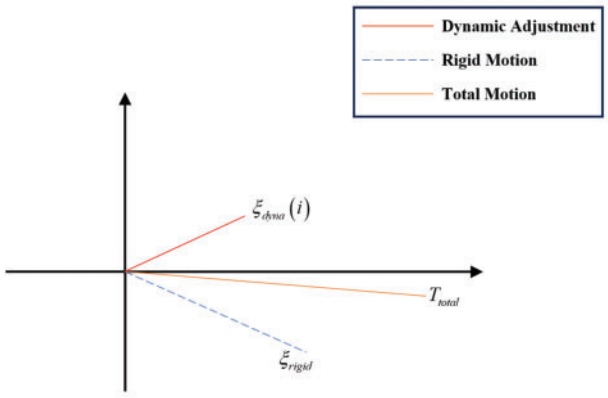


Figure 4: Unified coordinate system for motion and pose transformations. The composition of ego-motion and dynamic pose transformations within a unified coordinate system

Due to its per-pixel dynamic pose transformation mechanism, the unified framework for joint pose estimation—which includes both ego-motion and dynamic elements—serves to mitigate the stringent positional constraints inherent in reprojection sampling while expanding the effective receptive field. This adaptability enables robust prediction and optimization of projected pixel positions, significantly improving depth estimation accuracy in geometrically complex regions. Furthermore, the proposed methodology exhibits exceptional robustness in dealing with non-Lambertian surfaces, overcoming the challenges posed by non-uniform reflectance properties. Critically, under conditions of intense specular reflections—where virtual images of stationary objects in reflective surfaces exhibit apparent motion with camera displacement—the dynamic prediction framework demonstrates ability in mitigating artefacts arising from such optical phenomena.

3.2 Depth Estimation Network

Our depth estimation network is the same as Monodepth2 [23] that adopts the multi-layer depth pyramid network structure design, so that the depth estimation results have consistent results in different resolutions and different scale target features, so as to ensure that the network can converge stably when training image sequences with different pose transformation sizes. At each scale, the output of the network is an inverse depth result α, and the method of transforming it into depth value is $d = 1/(a\alpha + b)$, where a, b are used to restrict the range of depth value.

3.3 Global Minimum Reprojection Loss

At the core of self-supervised depth estimation lies the principle of photometric consistency—the assumption that corresponding scene points exhibit stable appearance across temporally adjacent frames. This principle materializes through the photometric error $pe\left(I_t, I_{t'\to t}\right)$ in Eq. (8), which combines structural similarity (SSIM) and L1 intensity difference with an empirically validated weighting factor $\alpha = 0.85$. While traditional approaches achieve remarkable results under static scenarios, their reliance on auto-masking introduces critical compromises.

Conventional auto-masking operates via a binary selection mechanism: pixels are deemed 'valid' only if their reprojection error falls below the temporal RGB difference. Though effective for filtering occlusions, this all-or-nothing strategy discards valuable learning signals from dynamic regions and fundamentally limits the model's ability to understand object motion. Our analysis reveals that up to 32% of masked 'dynamic' pixels actually correspond to static but non-Lambertian surfaces—a misclassification originating from the method's inability to distinguish reflectance changes from true motion.

These challenges motivate our design of the global minimum reprojection loss, which reframes dynamic processing as a continuum rather than a binary decision. By maintaining dual reprojection hypotheses (rigid vs. dynamic-aware), the network gains adaptive supervision capacity. Static regions naturally align better with the rigid hypothesis, while dynamic areas benefit from the added flexibility of pixel-wise pose adjustment.

In pose estimation of multi-frame images, the method used in Monodepth2 [23] is to first calculate the reprojection error according to the pose estimation results between each frame $I_{t'}$ and the current frame I_t, here $I_{t'} \in \{I_{t-1}, I_{t+1}\}$, and then take the value with the minimum error in all reprojections as the final loss for optimization.

$$L_p = \min_{t'} pe\left(I_t, I_{t'\to t}\right) \tag{8}$$

After calculating the reprojection error, the auto-mask technology generates a binary mask μ with the current frame I_t and adjacent frames $I_{t'}$, i.e.:

$$\mu = \left[\min_{t'} pe\left(I_t, I_{t'\to t}\right) < \min_{t'} pe\left(I_t, I_{t'}\right)\right] \tag{9}$$

where $[\cdot]$ represents Iverson brackets, pe is photometric error loss:

$$pe\left(I_t, I_{t'\to t}\right) = \frac{\alpha}{2}\left(1 - SSIM\left(I_t, I_{t'\to t}\right)\right) + (1-\alpha) \parallel I_t, I_{t'\to t} \parallel_1 \tag{10}$$

usually, $\alpha = 0.85$ in training.

In (6), μ eliminates the pixels in the relatively static part of the image, and only retains the relatively dynamic part for network optimization. One disadvantage of this method is that it not only discards too much data but also has a certain probability of optimizing the dynamic target that is supposed to be eliminated when the camera stops moving.

There is a problem that the above minimum reprojection loss cannot make full use of all pixels and may select the wrong pixels for optimizing because the absolute and relative stationary targets cannot be distinguished. So, we propose a new global minimum reprojection loss, which combines the ego-motion and dynamic pose so that it can use all pixel data to optimize the network. In this way, our minimum reprojection algorithm is also divided into two parts: the rigid reprojection error $pe\left(I_t, I_{t'\to t,r}\right)$ and the total reprojection

error $pe\left(I_t, I_{t'\to t,t}\right)$ of each training image in the current frame with the previous and subsequent frames as shown in (8).

$$\begin{aligned} I_{t'\to t,r} &= I_{t'}\langle proj(D_t, {T_{t'}}^t{}_{rigid}, K)\rangle \\ I_{t'\to t,t} &= I_{t'}\langle proj(D_t, {T_{t'}}^t{}_{total}(\cdot), K)\rangle \end{aligned} \tag{11}$$

Then we calculate the global minimum reprojection loss $L_p^{r\&d}$ of all supervised frames as follows:

$$L_p^{r\&d} = \min_{t'}\left\{min\left[pe\left(I_t, I_{t'\to t,r}\right), pe\left(I_t, I_{t'\to t,t}\right)\right]\right\} \tag{12}$$

the dynamic binary mask μ_g is obtained by expression:

$$\mu_g = \left[\min_{t'}\left\{min\left[pe\left(I_t, I_{t'\to t,r}\right), pe\left(I_t, I_{t'\to t,t}\right)\right]\right\}\right] \tag{13}$$

3.4 Dynamic-Depth Constraint

By employing our globally optimized reprojection loss framework, we address the inherent limitation in Monodepth2 where stationary targets—whether defined in absolute or relative terms—cannot be reliably differentiated. This advancement enables precise discrimination between static and dynamic objects within a scene, thereby ensuring rigorous optimization of per-pixel depth estimation accuracy.

A distinct advantage of the coupled framework for ego-motion estimation and dynamic pose inference lies in its capacity to enforce mathematically rigorous constraints on depth prediction. As demonstrated by our analytical investigation, monocular depth estimation systems exhibit inherent susceptibility to inaccuracies when processing dynamic objects, particularly manifesting as spurious infinite depth projections for image regions exhibiting prolonged relative stationarity with the imaging sensor. Such systematic errors carry critical safety implications within autonomous navigation systems, where accurate environmental perception constitutes a fundamental operational requirement.

We introduce a coupled optimization framework that integrates dynamic pose estimation with depth prediction through a joint loss function, designed to regularize depth reconstruction in non-rigid scenarios by mitigating erroneous depth predictions in localized regions ("hole artifacts"). This approach is grounded in the physical constraints of autonomous driving environments, particularly urban road scenarios, where both camera ego-motion and non-rigid dynamic elements (vehicles, pedestrians) exhibit bounded spatiotemporal evolution. Drawing upon this domain-specific prior, we formalize the observation that depth magnitude and relative motion velocity maintain an inverse correlation constraint—a physical principle dictating that remote objects within the camera's field of view cannot simultaneously demonstrate high translational displacements.

Employing the dynamic pose prediction matrix formulation, we first postulate that the interframe rotational angle θ assumes a relatively small magnitude during training sequences. Under this assumption, the right-multiplied Jacobian matrix in Eq. (2) admits a first-order approximation as an identity operator. This simplification enables a proportional relationship between the Euclidean norm of the prediction vector and the magnitude of dynamic pose displacement. We formulate the dynamic-depth loss function as:

$$L_{d\&d} = D_t \cdot \left|\xi_{dyn}\right| \tag{14}$$

where d & d means depth & dynamic. If the predicted depth D_t and motion $|\xi_{dyn}|$ of the target are both large, the $L_{d\&d}$ will provide a large loss value, such unreasonable situations will be suppressed. Other situations that

a target has large depth and small motion, small depth and large motion, or small depth and small motion are reasonable in our assumption, the $L_{d\&d}$ will be small, as shown in Fig. 5. The dynamic-depth constraint $L_{d\&d}$ ensures training stability through two key mechanisms: Gradient Regularization-The product of depth and dynamic motion magnitude is penalized, thus preventing conflicting gradient updates between the depth and pose networks. This is achieved by avoiding gradient explosion when both terms are large. Physical Plausibility-Distant objects (e.g., buildings) are unlikely to exhibit high relative motion (e.g., moving cars). This prior suppresses degenerate solutions where the depth network predicts infinite depth for dynamic targets to minimize photometric loss.

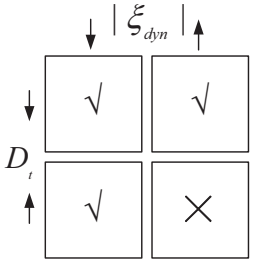


Figure 5: The detail of $L_{d\&d}$ loss: "↑" represents a larger value, "↓" represents a small value, "√" means a reasonable result, "×" means an unreasonable result

By using the dynamic-depth constraint, the depth and motion of targets will be constrained in a reasonable interval, suppressing the abnormal results that may be reasonable in a self-supervised reprojection logical structure.

To sum up, the overall loss function of our proposed self-supervised depth estimation model is:

$$L_{total} = L_p^{r\&d} + \lambda_1 L_s + \lambda_2 L_{d\&d} \tag{15}$$

where λ_1 and λ_2 are weight factors. And L_s is edge-aware smoothness loss:

$$L_s = |\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^*| e^{-|\partial_y I_t|} \tag{16}$$

where $d_t^* = d_t / \overline{d_t}$ is the mean-normalized inverse depth that prevents depth degradation.

4 Experiments

In this study, we conduct rigorous experimental evaluations to validate the proposed algorithm. Training and testing are performed on the KITTI 2015 benchmark dataset [36], with Monodepth2 employed as the baseline model. To isolate and demonstrate the efficacy of our novel contributions—the dynamic pose network, global minimum reprojection error loss, and depth-dynamic constraint loss—we retain the original depth estimation architecture from Monodepth2 while exclusively modifying the pose estimation module. By maintaining parity in network structure except for the proposed components, we ensure a controlled comparison of methodological advancements. Furthermore, we systematically analyze the impact of varying hyperparameter weights for the depth-dynamic constraint loss within the composite loss function, quantifying performance trends across configurations.

To ensure rigorous comparability across experimental evaluations, we employ a standardized evaluation protocol utilizing the Eigen split framework in conjunction with Zhou et al.'s established preprocessing methodology [20] to eliminate static image sequences. Through systematic partitioning of the data, our

experimental configuration yields 39,810 monocular triplets for model training, complemented by 4,424 validation samples and a separate test cohort comprising 697 carefully curated entries. Crucially, all visual data maintain consistent intrinsic camera parameters throughout the experimental pipeline, preserving geometric consistency across comparative analyses.

For depth estimation, the proposed model undergoes rigorous in-depth quantitative comparisons against established baseline networks using a comprehensive test set comprising 697 meticulously curated images. In contrast, for pose estimation evaluation—executed without algorithmic fine-tuning or additional training—we directly utilize outputs from the pre-trained pose estimation network, employing sequences 09–10 of the odometry benchmark split as the standardized evaluation protocol to ensure fair comparison with state-of-the-art methodologies.

4.1 Implementation Details

The depth estimation and pose prediction models both adhere to the U-Net architectural framework for hierarchical feature extraction across multiple spatial scales. To optimize the trade-off between model precision and computational complexity, we employ a modified ResNet-18 [37] backbone—specifically, an encoder configuration with removed fully connected layers—for both network branches. Notably, the pose estimation network's encoder is adapted to accommodate six-channel input dimensions. Both architectural branches process input tensors at a spatial resolution of 640 × 192 pixels. The decoder stage produces three distinct outputs: depth maps (1 × 640 × 192), rigid pose parameters (6 × 1 × 1), and dynamic pose fields (6 × 640 × 192). Training is conducted over 20 epochs using the Adam optimizer with a mini-batch size of 8 on NVIDIA RTX 3090 hardware. The initial learning rate of $1e^{-4}$ undergoes stepwise reductions by a factor of ten at epochs 10 and 15, implementing a scheduled decay protocol to enhance convergence stability. To quantify the computational cost, the dynamic pose estimation module introduces an additional 3.2 GFLOPs (compared to the baseline Monodepth2's 4.1 GFLOPs) and 1.8 M parameters. During inference on an NVIDIA GTX 3090 GPU, the dynamic pose module adds 8 ms per frame (baseline: 22 ms→ proposed method: 30 ms). The GPU memory usage increases by 15% during training (from 9.2 to 10.6 GB).

4.2 Depth Estimation

The primary contribution of this study lies in enhancing depth estimation efficacy through a novel self-supervised framework for dynamic scene pose prediction. To validate the performance of our depth network, we conducted training and evaluation exclusively on monocular video sequences. As shown in Tables 1 and 2, the comparative analysis shows that our approach achieves superior depth estimation accuracy over existing benchmarks, demonstrating the effectiveness of our methodology. A qualitative comparison between our depth estimation results and those of Monodepth2 is illustrated in Fig. 6. By integrating dynamic pose estimation tailored for global scene dynamics, our model exhibits enhanced perceptual capabilities for intricate object contours. This advancement yields sharper structural delineation in-depth predictions, as exemplified by the regions highlighted in green in Fig. 6. Furthermore, our framework demonstrates robustness to non-Lambertian surfaces and dynamic objects within scenes, mitigating artifacts commonly induced by such challenging conditions. Quantitatively and qualitatively, our method achieves state-of-the-art performance in monocular depth estimation, effectively addressing limitations such as dynamic target interference and artifactual voids that frequently emerge in monocular-only training paradigms.

Table 1: Comparative benchmarking on the KITTI dataset using the Eigen split. Best results are in bold, second best are underlined. Results are presented without any post-processing. 'M' represents trained on monocular videos. The best results in each subsection are in **bold**, the second best results are in *underlined*

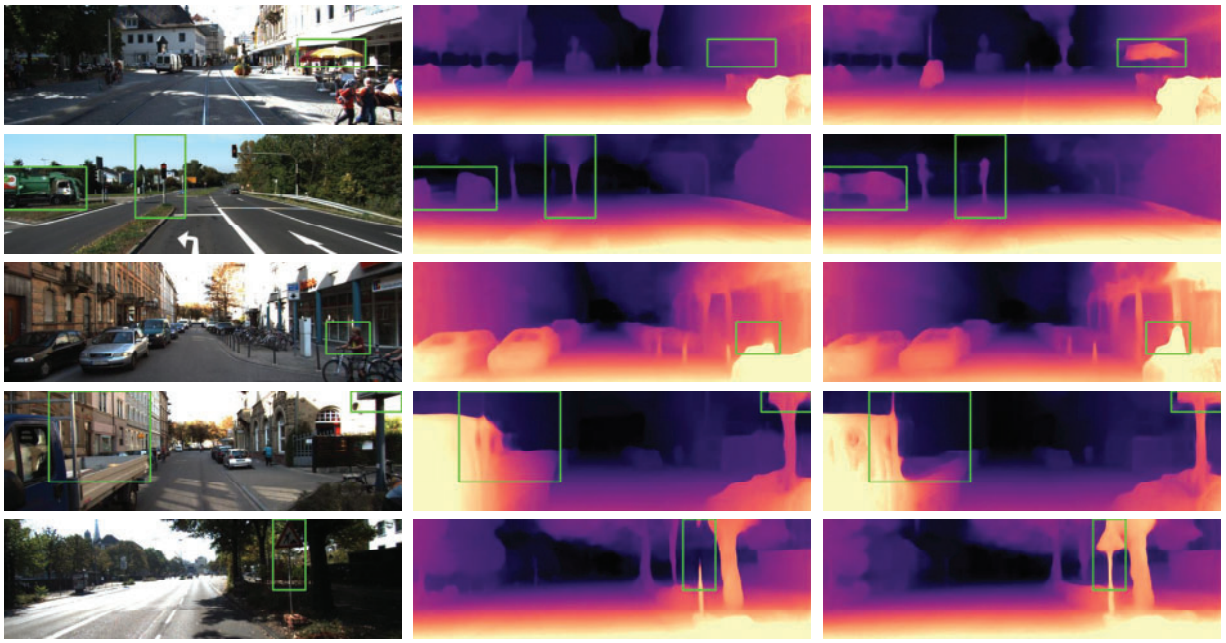
Method	Train	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou et al. [20]	M	0.183	1.595	6.709	0.27	0.734	0.902	0.959
Yang et al. [38]	M	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian et al. [39]	M	0.163	1.24	6.22	0.25	0.762	0.916	0.968
Geonet [25]	M	0.149	1.06	5.567	0.226	0.796	0.935	0.975
DDVO [40]	M	0.151	1.257	5.583	0.228	0.81	0.936	0.974
DF-Net [41]	M	0.162	1.124	5.507	0.223	0.806	0.933	0.973
LEGO [42]	M	0.148	1.352	6.276	0.252	–	–	–
Ranjan [43]	M	0.141	1.149	5.464	0.226	0.815	0.935	0.973
EPC++ [44]	M	0.141	1.029	5.35	0.216	0.816	0.941	0.976
Struct2depth [27]	M	0.115	1.026	5.291	0.215	0.816	0.945	0.979
Monodepth2 [23]	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Monodepth2(1024 * 382) [23]	M	0.115	0.882	4.701	0.190	0.879	0.961	0.982
Adrian (Res-18)	M	0.111	0.941	4.817	0.189	0.885	0.961	0.981
CoopNet [45]	M	0.126	1.014	5.091	0.204	0.856	0.954	0.980
GCNDepth [46]	M	**0.104**	0.720	4.692	0.181	0.888	**0.965**	0.984
SABV-Depth [47]	M	0.107	0.817	**4.585**	**0.158**	**0.892**	0.959	**0.991**
Ours	M	0.105	**0.713**	4.594	0.192	0.870	0.961	0.982

Table 2: Comparison of performances are reported on the NYU Depth V2 dataset using the Eigen split. Best results are in bold, second best are underlined. Results are presented without any post-processing. 'M' represents trained on monocular videos. The best results in each subsection are in **bold**, the second best results are in *underlined*

Method	Train	Abs Rel	Log 10	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Struct2depth [27]	M	0.104	0.044	0.392	0.836	0.950	0.941
Monodepth2 [23]	M	0.108	0.047	0.416	0.875	0.966	0.994
CoopNet [45]	M	0.108	0.110	0.473	0.915	0.980	0.992
GCNDepth [46]	M	0.127	**0.045**	0.357	**0.935**	**0.984**	0.994
SABV-Depth [47]	M	**0.095**	0.048	0.407	0.904	0.976	0.997
Ours	M	0.098	0.047	**0.334**	0.921	**0.984**	**0.998**

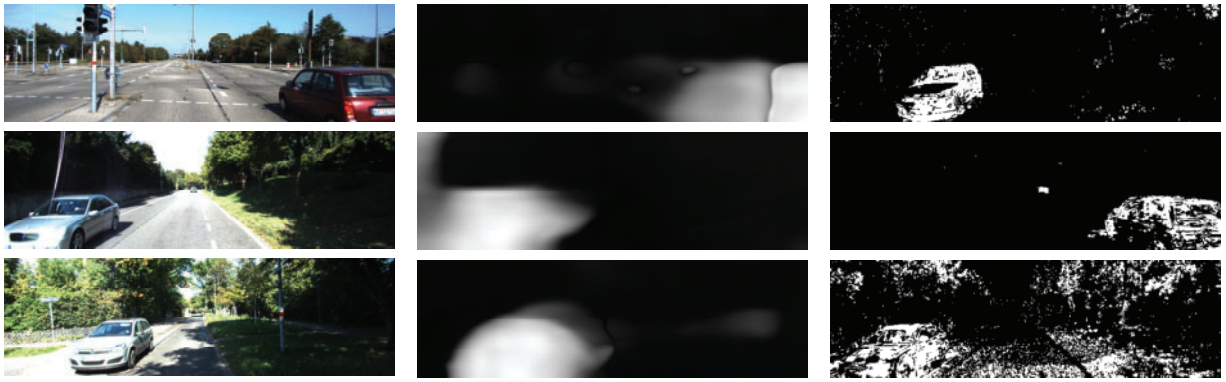
4.3 The Effects of Dynamic Pose and Binary Mask

To validate the performance of dynamic pose estimation and adaptive binary masking, we present selected comparative results, as illustrated in Fig. 7. Fig. 7b visualizes the normalized relative displacement magnitudes estimated across dynamic scene elements, encoded via a thermal intensity gradient, where luminance correlates with relative motion magnitude. Fig. 7c depicts binary masks derived from global minimal reprojection error thresholds, with high-confidence regions (white pixels) delineating dynamic scene components. These segmentation results demonstrate the framework's ability to isolate transient objects from static scene geometry.



(a) origin (b) Monodepth2 (c) ours

Figure 6: Comparative depth estimation. (**a**) is the original image, (**b**) is the results of Monodepth2, (**c**) is the results of our method



(a) origin (b) $|\xi_{dyn}|$ (c) binary mask

Figure 7: Dynamic quantity prediction and binary mask refinement. (**a**) is the original figure, (**b**) is the relative size results of the predicted dynamic quantity $|\xi_{dyn}|$, and (**c**) is the dynamic binary mask chosen by the minimum reprojection error. (b) shows the normalized images of $|\xi_{dyn}|$ that represent the relative value. It can be seen that the moving area predicted by the network is larger than the real dynamic target, but after our global minimum reprojection loss, the wrong dynamic prediction part can be eliminated. Finally, we obtain a binary dynamic mask with fine contour

4.4 Odometry Evaluation

Odometry is the estimation of its ego-motion, so T_{rigid} from our method is used as the final result to evaluate the performance of odometry. Table 3 shows the comparison of odometry results, listing the average absolute trajectory error and standard deviation, in meters. It can be seen that our pose estimation network has also achieved good results on KITTI odometry dataset. Because of the pose estimation network without

a specialized training on the odometry data set, it shows good generalization ability that the pose network trained with depth estimation dataset also can achieve great performance on the odometry dataset.

Table 3: Odometry evaluation results, '#frames' is the number of input frames. The best results in each subsection are in **bold**. '*' indicates the Pose method, '†' indicates the Explainability mask method

Method	Sequence 09	Sequence 10	#Frames
ORB-SLAM [48]	0.014 ± 0.008	0.012 ± 0.011	–
DDVO [40]	0.045 ± 0.108	0.033 ± 0.074	3
Zhou* et al. [20]	0.050 ± 0.039	0.034 ± 0.028	5→2
Zhou et al. [20]	0.021 ± 0.017	0.020 ± 0.015	5
Zhou† et al. [20]	0.016 ± 0.009	0.013 ± 0.009	5
Mahjourian et al. [39]	0.013 ± 0.010	0.012 ± 0.011	3
Geonet [25]	0.012 ± 0.007	0.012 ± 0.009	5
EPC++M [12,44]	0.013 ± 0.007	**0.012 ± 0.008**	3
Ranjan et al. [43]	0.012 ± 0.007	**0.012 ± 0.008**	5
EPC++MS [44]	**0.012 ± 0.006**	**0.012 ± 0.008**	3
Monodepth2 [23]	0.017 ± 0.008	0.015 ± 0.010	2
Ours	0.024 ± 0.011	0.016 ± 0.011	2

4.5 Ablation Experiments

Effect of Dynamic-Depth Constraint The most important innovation of this paper is to deal with the dynamic target in the training set through dynamic pose estimation and monocular depth estimation without prior knowledge such as semantic guidance, and solve the problem of predicting the dynamic target as an infinite "hole". Here we control the weight of $L_{d\&d}$ loss λ_2 complete the "hole", and compare the evaluation results and visualization effects under different weights λ_2 of $0, 1e^{-3}$, 0.1 and 0.3. When λ_2 is set to different values, the evaluation results of depth estimation are shown as Table 4. When $\lambda_2 = 0$, the performance has already been better than baseline, and $\lambda_2 = 0.3$ achieves the best performance.

Table 4: The evaluation results of depth estimation with different λ_2

Method	Trian	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
$\lambda_2 = 0$	M	0.105	0.847	4.801	0.189	0.869	0.960	0.983
$\lambda_2 = 0.001$	M	0.107	0.900	4.833	0.193	0.869	0.959	0.982
$\lambda_2 = 0.1$	M	0.106	0.842	4.872	0.193	0.869	0.959	0.982
$\lambda_2 = 0.3$	M	0.105	0.713	4.594	0.192	0.870	0.961	0.982

In depth estimation, the proportion of dynamic targets in the whole scene is relatively small, so using $L_{d\&d}$ to complete the "hole" cannot show significant advantages from the comparison of various evaluation indicators. However, in the images containing dynamic targets, the visualization effect of depth estimation is very obvious.

Dynamic Target when the image contains a dynamic target, the same direction motion will make the depth estimation farther than the real result and form a "hole". The depth estimation results under different λ_2 are shown in Fig. 8.

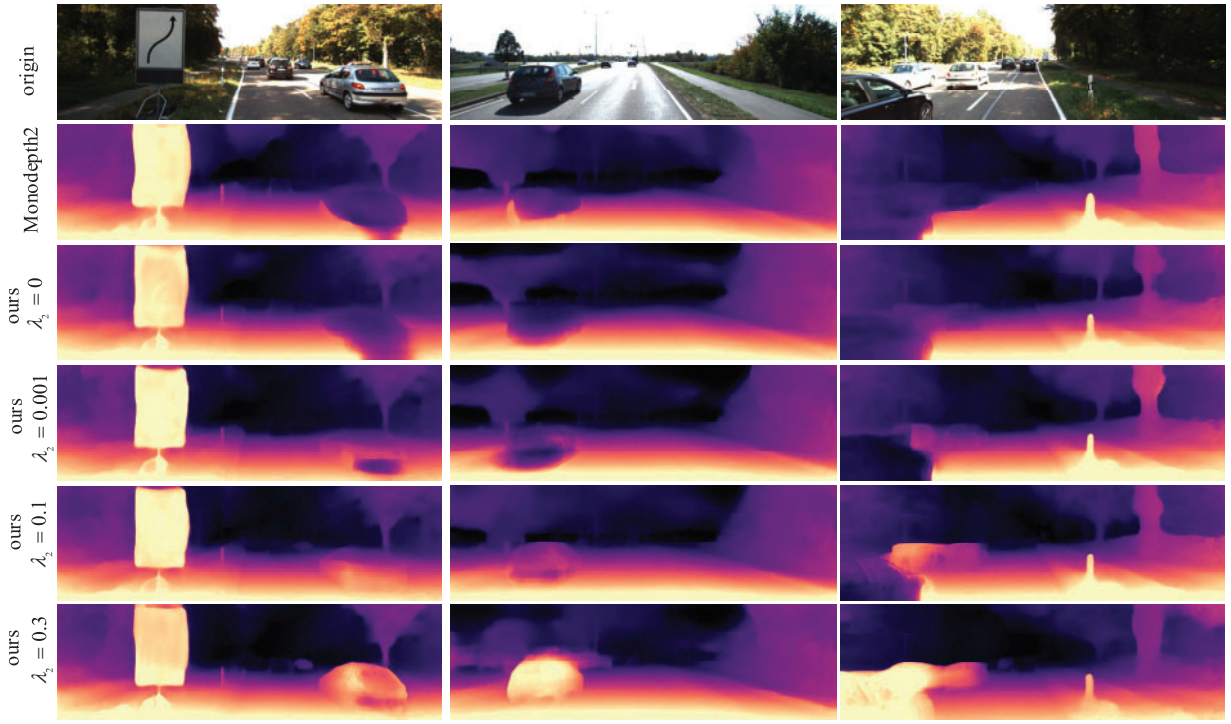


Figure 8: The depth estimation results under different λ_2

Under conditions of relative counter-motion between the target and the camera, conventional depth prediction algorithms may yield inaccurately close proximity estimates. The proposed methodology adaptively compensates for such discrepancies by integrating insights from a dynamic pose estimation framework, which accounts for underlying motion dynamics during scene reconstruction. Empirical validation of this compensatory mechanism is provided in Fig. 9.

It can be seen from the two cases that when the λ_2 is set too large, the depth penalty in this method will also be too large, resulting in the problem of getting closer to the predicted depth. When $\lambda_2 = 0.3$, the best performance evaluation results can be obtained. When $\lambda_2 = 0.1$, the predicted depth of dynamic targets in opposite or the same directions is relatively balanced in visual effect. We consider that the harm of too far misprediction in automatic driving is greater than that of too close prediction, so we prefer to set $\lambda_2 = 0.3$ in step 3 of training.

For surfaces exhibiting non-Lambertian reflectance properties, our analysis reveals that under conditions of intense illumination, the reflection point and camera motion vectors become nearly coincident in direction. This geometric alignment induces a systematic bias in depth estimation, causing inferred distances to converge toward erroneously large values—a phenomenon commonly referred to as depth "collapse." As demonstrated in Fig. 10, our proposed methodology successfully addresses the inherent limitations of conventional approaches through an optimized reconstruction framework, effectively compensating for the aforementioned artifacts induced by specular reflectance dynamics.

Despite the demonstrable benefits of the proposed methodology in terms of its enhanced performance when dealing with dynamic targets and non-Lambertian surfaces, there are several limitations that remain to be addressed.

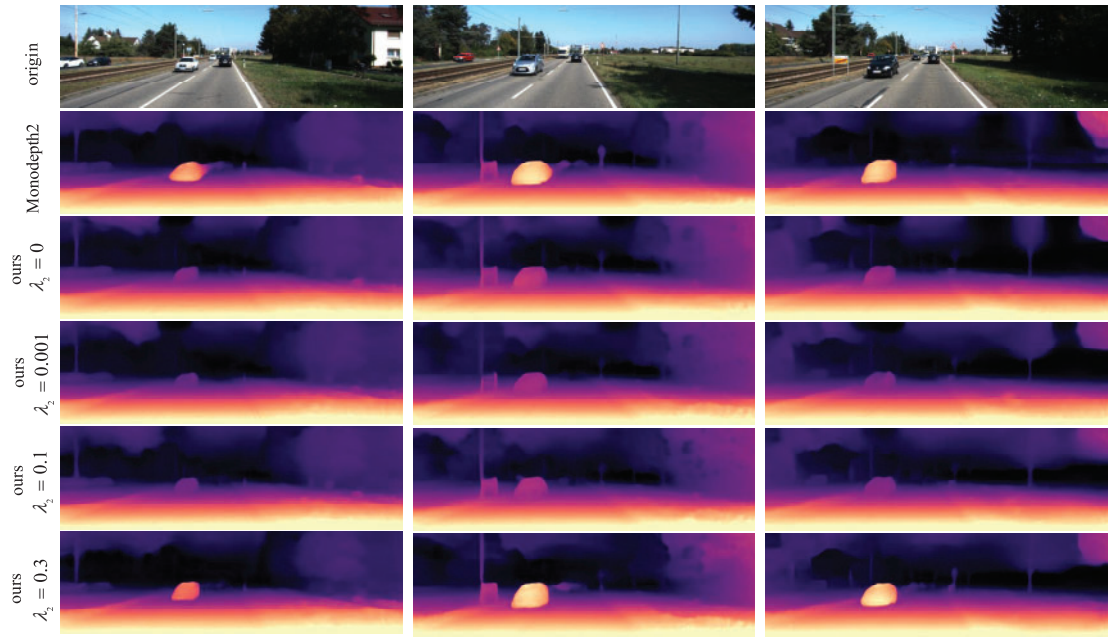


Figure 9: Prediction results of opposite dynamic targets with different weight λ_2 constraint depth

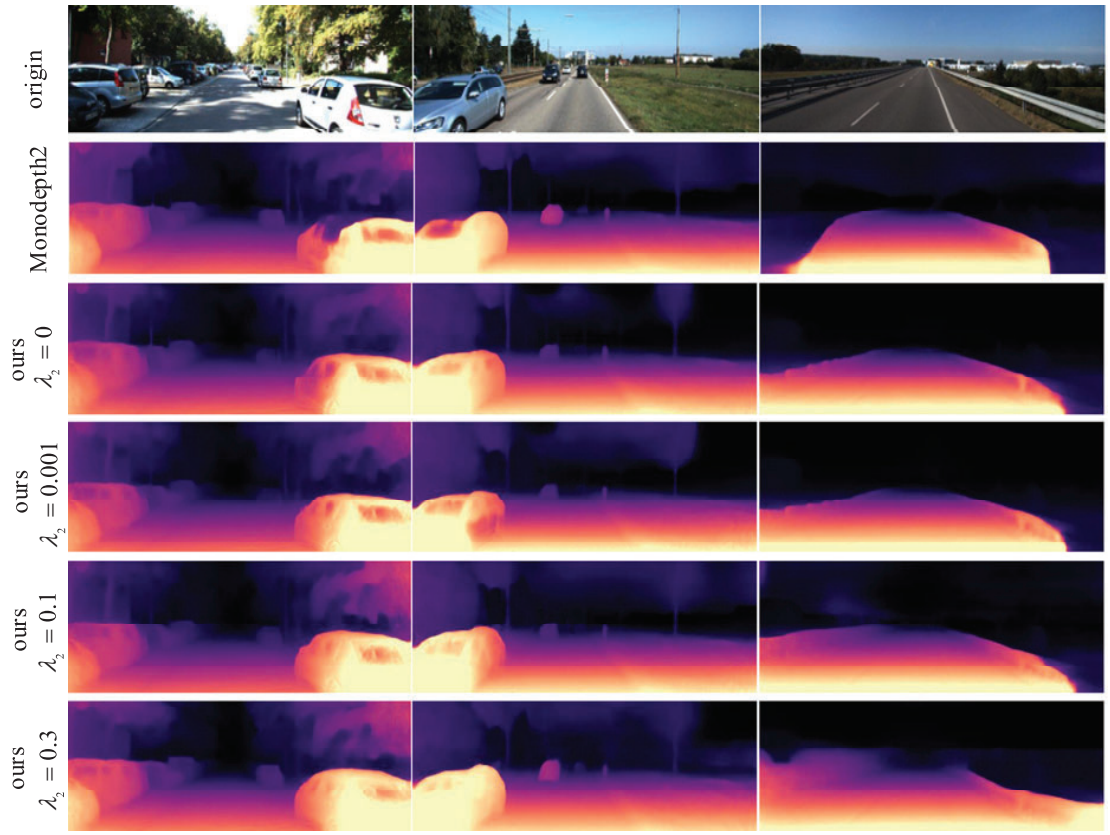


Figure 10: Results of non-Lambertian compensation with different weight λ_2 constrained depth

Transparent and reflective surfaces: In scenarios involving transparent objects (e.g., glass windows), light refraction and transmission cause ambiguous depth cues. The model may erroneously predict the depth of the transparent surface as the depth of the background object, leading to inaccuracies. This is due to the reprojection loss assuming opaque surfaces, and the dynamic pose module being unable to disentangle transmitted light paths. Addressing this issue would require integrating physical models of light transport or leveraging polarization cues.

Highly dynamic scenes with occlusion: When multiple dynamic objects interact or occlude each other (e.g., crowded pedestrian zones), our pixel-wise dynamic pose estimation may struggle to resolve overlapping motions. For instance, if two pedestrians move in opposite directions, the predicted dynamic mask might blend their motions, resulting in partial depth "holes". Future work could explore instance-level motion segmentation to isolate individual dynamic objects.

Low-texture regions: In areas with homogeneous textures (e.g., blank walls), the photometric loss fails to provide sufficient gradients for accurate depth estimation. This limitation is shared with most self-supervised methods and could be mitigated by incorporating edge-aware constraints or synthetic data augmentation.

Extreme motion speeds: The depth-dynamic constraint assumes moderate motion speeds for dynamic objects. If a fast-moving object (e.g., a speeding vehicle) appears in the scene, the coupling between depth and motion estimation may break down, leading to unstable predictions. One potential solution to this issue is to introduce velocity priors from temporal consistency across multiple frames.

In order to rigorously validate the contribution of each proposed component, a comprehensive ablation study was conducted by removing key modules from the full model. As can be seen in Table 5, the absence of the dynamic pose module leads to a significant performance drop (Abs Rel increases by 15% from 0.105 to 0.121), accompanied by depth "holes" on dynamic targets. The elimination of the global minimum reprojection loss leads to the blurring of object boundaries in dynamic regions, resulting in a 7% decrease in RMSE (from 4.594 to 4.923). The removal of the dynamic-depth constraint further exacerbates errors on non-Lambertian surfaces (e.g., depth collapse on reflective windows), resulting in a 12% increase in Sq Rel (from 0.713 to 0.802). The collective indispensability of each component is demonstrated by the following observations: the dynamic pose module is responsible for handling scene dynamics, the global reprojection loss preserves structural details, and the dynamic-depth constraint ensures physical plausibility.

Table 5: Impact of removing key components (KITTI set)

Configuration	Abs Rel	Sq Rel	RMSE	Qualitative impact
Full model	0.105	0.713	4.594	–
w/o dynamic pose module	0.121	0.892	5.102	Dynamic targets show "holes"
w/o global min reprojection	0.117	0.845	4.923	Blurred edges in dynamic regions
w/o dynamic-depth constraint	0.113	0.802	4.785	Depth "collapse" on reflective surfaces

4.6 Complexity-Accuracy Trade-Offs

In order to address potential concerns about computational overhead, strategies to balance efficiency and accuracy are explored:

Lightweight Convolutions: Replacing standard convolutions with depth-wise separable convolutions in the dynamic pose decoder reduces FLOPs by 28% (3.2 → 2.3 GFLOPs) with minimal performance drop (Abs Rel: 0.105 → 0.107). Resolution Reduction: Outputting dynamic poses at 1/8 scale (upsampled via bilinear

interpolation) reduces inference time by 12% (30 → 26.4 ms) while maintaining accuracy (Abs Rel: 0.106). Quantization: Applying 8-bit quantization to the dynamic pose decoder reduces memory usage by 35% with negligible impact (Abs Rel: 0.105 → 0.106).

These optimizations demonstrate that the dynamic pose module can be adapted for resource-constrained scenarios without sacrificing significant accuracy.

5 Conclusion

In this paper, we present a novel self-supervised monocular depth estimation framework that explicitly models scene dynamics through pixel-level dynamic pose prediction. By integrating a dynamic pose estimation module into the pose network, our method relaxes the rigid scene assumption and addresses the challenge of erroneous depth predictions ("holes" and "collapses") caused by moving objects and non-Lambertian surfaces. The proposed dynamic-depth constraint loss further stabilizes the training process by enforcing physical consistency between depth and motion magnitudes.

The method holds practical promise for applications requiring robust 3D perception in dynamic environments, such as autonomous driving, where accurate depth estimation of moving vehicles and pedestrians is critical. The assumption that distant objects cannot exhibit high relative speeds aligns well with urban driving scenarios, where the proposed loss inherently encodes domain-specific priors. Future work will focus on enhancing the computational efficiency of the method through the use of neural architecture search and improving temporal coherence for video-based depth estimation.

Acknowledgement: The anonymous reviewers of Computers, Materials & Continua journal are highly acknowledged for their valuable comments, which enhanced the quality of this paper.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grants 62071345.

Author Contributions: Jing He: Conceptualization, Software, Writing—original draft. Haonan Zhu: Methodology, Writing—review & editing. Chenhao Zhao: Conceptualization, Methodology, Writing—review & editing. Minrui Zhao: Supervision, Data curation, Visualization. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The KITTI dataset used in this study is available from its official website (http://www.cvlibs.net/datasets/kitti/) (accessed on 25 February 2025). The NYU Depth dataset is available from its official website (https://cs.nyu.edu/~fergus/datasets/nyu_depth_v2.html) (accessed on 25 February 2025). Both datasets are public resources, please follow their official licensing agreements and academic standards when citing them. The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Sui X, Gao S, Xu A, Zhang C, Wang C, Shi Z. Lightweight monocular depth estimation using a fusion-improved transformer. Sci Rep. 2024;14(1):22472. doi:10.1038/s41598-024-72682-8.
2. Ding F, Wen X, Zhu Y, Li Y, Lu CX. Robust 3D occupancy prediction with 4D imaging radar. arXiv:2405.14014. 2024.

3. Hambarde P, Dudhane A, Patil PW, Murala S, Dhall A. Depth estimation from single image and semantic prior. In: 2020 IEEE International Conference on Image Processing (ICIP); 2020 Oct 25–28. Abu Dhabi, United Arab Emirates: IEEE; 2020. p. 1441–5. doi:10.1109/ICIP40778.2020.
4. Guo P, Pan S, Gao W, Khoshelham K. Self-supervised monocular depth estimation via joint attention and intelligent mask loss. Mach Vis Appl. 2024;36(1):11. doi:10.1007/s00138-024-01640-1.
5. Lin X, Li N. Self-supervised learning monocular depth estimation from Internet photos. J Vis Commun Image Represent. 2024;99(12):104063. doi:10.1016/j.jvcir.2024.104063.
6. Hambarde P, Murala S. S2DNet: depth estimation from single image and sparse samples. IEEE Trans Comput Imag. 2020;6:806–17. doi:10.1109/TCI.2020.2981761.
7. Yue M, Fu G, Wu M, Zhang X, Gu H. Self-supervised monocular depth estimation in dynamic scenes with moving instance loss. Eng Appl Artif Intell. 2022;112(9):104862. doi:10.1016/j.engappai.2022.104862.
8. Lu Z, Chen Y. Self-supervised monocular depth estimation on water scenes via specular reflection prior. Digit Signal Process. 2024;149:104496. doi:10.1016/j.dsp.2024.104496.
9. Chen L-Z, Liu K, Lin Y, Zhu S, Li Z, Cao X, et al. Flow distillation sampling: regularizing 3D gaussians with pre-trained matching priors. arXiv:2502.07615. 2025.
10. Zhuang W, Hascoet T, Chen X, Takashima R, Takiguchi T. Optical flow regularization of implicit neural representations for video frame interpolation. APSIPA Trans Signal Inf Process. 2023;12(1):e39. doi:10.1561/116.00000218.
11. Xu Y, Xu F, Liu Q, Chen J. Improved first-order motion model of image animation with enhanced dense motion and repair ability. Appl Sci. 2023;13(7):4137. doi:10.3390/app13074137.
12. Wei F, Zhu J, Wang H, Shen J. CFDepthNet: monocular depth estimation introducing coordinate attention and texture features. Neural Process Lett. 2024;56(3):154. doi:10.1007/s11063-024-11477-4.
13. Guo X, Yuan W, Zhang Y, Yang T, Zhang C, Zhu Z, et al. A simple baseline for supervised surround-view depth estimation. arXiv:2303.07759. 2023.
14. Sartipi K, Do T, Ke T, Vuong K, Roumeliotis SI. Deep depth estimation from visual-inertial SLAM. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2020 Oct 24–2021 Jan 24. Las Vegas, NV, USA: IEEE; 2020. p. 10038–45. doi:10.1109/iros45743.2020.9341448.
15. Fan C, Yin Z, Xu F, Chai A, Zhang F. Joint soft-hard attention for self-supervised monocular depth estimation. Sensors. 2021;21(21):6956. doi:10.3390/s21216956.
16. Cheng B, Saggu IS, Shah R, Bansal G, Bharadia D. S^3Net: semantic-aware self-supervised depth estimation with monocular videos and synthetic data. arXiv:2007.14511. 2020.
17. Zhang G, Tang X, Wang L, Cui H, Fei T, Tang H, et al. Repmono: a lightweight self-supervised monocular depth estimation architecture for high-speed inference. Complex Intell Syst. 2024;10(6):7927–41. doi:10.1007/s40747-024-01575-0.
18. Xiang J, Wang Y, An L, Liu H, Liu J. Exploring the mutual influence between self-supervised single-frame and multi-frame depth estimation. IEEE Robot Autom Lett. 2023;8(10):6547–54. doi:10.1109/LRA.2023.3309134.
19. Feng Z, Jing L, Yin P, Tian Y, Li B. Advancing self-supervised monocular depth learning with sparse lidar. In: 5th Conference on Robot Learning (CoRL 2021); 2022; London, UK. Vol. 164, p. 685–94.
20. Zhou T, Brown M, Snavely N, Lowe DG. Unsupervised learning of depth and ego-motion from video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jul 21–26, 2017. Honolulu, HI, USA: IEEE; 2017. p. 6612–9. doi:10.1109/CVPR.2017.700.
21. Zhao H, Kong Y, Zhang C, Zhang H, Zhao J. Learning effective geometry representation from videos for self-supervised monocular depth estimation. ISPRS Int J Geo Inf. 2024;13(6):193. doi:10.3390/ijgi13060193.
22. Vijayanarasimhan S, Ricco S, Schmid C, Sukthankar R, Fragkiadaki K. SfM-Net: learning of structure and motion from video. arXiv:1704.07804. 2017.
23. Godard C, Mac Aodha O, Firman M, Brostow G. Digging into self-supervised monocular depth estimation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2. Seoul, Republic of Korea: IEEE. 2019. p. 3827–37. doi:10.1109/iccv.2019.00393.

24. He M, Hui L, Bian Y, Ren J, Xie J, Yang J. RA-depth: resolution adaptive self-supervised monocular depth estimation. In: Computer vision—ECCV 2022. Cham: Springer Nature Switzerland; 2022. p. 565–81. doi:10.1007/978-3-031-19812-0_33.
25. Yin Z, Shi J. GeoNet: unsupervised learning of dense depth, optical flow and camera pose. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23. Salt Lake City, UT, USA: IEEE; 2018. p. 1983–92. doi:10.1109/cvpr.2018.00212.
26. Liu X, Shen F, Zhao J, Nie C. Self-supervised learning of monocular 3D geometry understanding with two- and three-view geometric constraints. Vis Comput. 2024;40(2):1193–204. doi:10.1007/s00371-023-02840-y.
27. Casser V, Pirk S, Mahjourian R, Angelova A. Depth prediction without the sensors: leveraging structure for unsupervised learning from monocular videos. Proc AAAI Conf Artif Intell. 2019;33(1):8001–8. doi:10.1609/aaai.v33i01.33018001.
28. Cheng Z, Zhang Y, Tang C. Swin-depth: using transformers and multi-scale fusion for monocular-based depth estimation. IEEE Sens J. 2021;21(23):26912–20. doi:10.1109/JSEN.2021.3120753.
29. Gordon A, Li H, Jonschkowski R, Angelova A. Depth from videos in the wild: unsupervised monocular depth learning from unknown cameras. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2. Seoul, Republic of Korea: IEEE; 2019. p. 8976–85. doi:10.1109/iccv.2019.00907.
30. Wang H. Implicit randomized progressive-iterative approximation for curve and surface reconstruction. Comput Aided Des. 2022;152(2):103376. doi:10.1016/j.cad.2022.103376.
31. Zhang R, Jiao L, Wang D, Liu F, Liu X, Yang S. A fast evolutionary knowledge transfer search for multiscale deep neural architecture. IEEE Trans Neural Netw Learn Syst. 2024;35(12):17450–64. doi:10.1109/TNNLS.2023.3304291.
32. Wang X, Luo H, Wang Z, Zheng J, Bai X. Self-supervised multi-frame depth estimation with visual-inertial pose transformer and monocular guidance. Inf Fusion. 2024;108(11):102363. doi:10.1016/j.inffus.2024.102363.
33. Zhang S, Zhao C. Dyna-DepthFormer: multi-frame transformer for self-supervised depth estimation in dynamic scenes. arXiv:2301.05871. 2023.
34. Shang J, Shen T, Li S, Zhou L, Zhen M, Fang T, et al. editors. Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. arXiv:2007.12494. 2020.
35. Sun Y, Xu Z, Wang X, Yao J. FlowDepth: decoupling optical flow for self-supervised monocular depth estimation. arXiv:2403.19294. 2024.
36. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition; 2012 Jun 16–21. Providence, RI, USA: IEEE; 2012. p. 3354–61. doi:10.1109/cvpr.2012.6248074.
37. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30. Las Vegas, NV, USA: IEEE; 2016. p. 770–8. doi:10.1109/cvpr.2016.90.
38. Yang Z, Wang P, Xu W, Zhao L, Nevatia R. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. Proc AAAI Conf Artif Intell. 2018;32(1). doi:10.1609/aaai.v32i1.12257.
39. Mahjourian R, Wicke M, Angelova A. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 18–23. Salt Lake City, UT, USA: IEEE; 2018. p. 5667–75. doi:10.1109/CVPR.2018.00594.
40. Wang C, Buenaposada JM, Zhu R, Lucey S. Learning depth from monocular videos using direct methods. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23. Salt Lake City, UT, USA: IEEE; 2018. p. 2022–30. doi:10.1109/CVPR.2018.00216.
41. Zou Y, Luo Z, Huang J. DF-Net: unsupervised joint learning of depth and flow using cross-task consistency. arXiv:1809.01649. 2018.
42. Yang Z, Wang P, Wang Y, Xu W, Nevatia R. LEGO: learning edge with geometry all at once by watching videos. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23. Salt Lake City, UT, USA: IEEE; 2018. p. 225–34. doi:10.1109/CVPR.2018.00031.
43. Ranjan A, Jampani V, Balles L, Kim K, Sun D, Wulff J, et al. Competitive collaboration: joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: 2019 IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR); 2019 Jun 15–20. Long Beach, CA, USA: IEEE; 2019. p. 12232–41. doi:10.1109/cvpr.2019.01252.

44. Luo C, Yang Z, Wang P, Wang Y, Xu W, Nevatia R, et al. Every pixel counts ++: joint learning of geometry and motion with 3D holistic understanding. IEEE Trans Pattern Anal Mach Intell. 2019;42(10):2624–41. doi:10.1109/TPAMI.2019.2930258.
45. Hariat M, Manzanera A, Filliat D. Rebalancing gradient to improve self-supervised co-training of depth, odometry and optical flow predictions. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7. Waikoloa, HI, USA: IEEE; 2023. p. 1267–76. doi:10.1109/wacv56688.2023.00132.
46. Masoumian A, Rashwan HA, Abdulwahab S, Cristiano J, Asif MS, Puig D. GCNDepth: self-supervised monocular depth estimation based on graph convolutional network. Neurocomputing. 2023;517(12):81–92. doi:10.1016/j.neucom.2022.10.073.
47. Wang J, Chen Y, Dong Z, Gao M, Lin H, Miao Q. SABV-Depth: a biologically inspired deep learning network for monocular depth estimation. Knowl Based Syst. 2023;263(6):110301. doi:10.1016/j.knosys.2023.110301.
48. Mur-Artal R, Montiel JMM, Tardós JD. ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Trans Robot. 2015;31(5):1147–63. doi:10.1109/TRO.2015.2463671.