VFM-Dynamo: Accurate Non-rigid Motion Identification via Vision Foundation Model for Self-supervised Monocular Depth Estimation

Qianqian Du[1,2], **Hui Yin**[1,2,5*], **Xingyu Miao**[4] and **Zhengyin Liang**[3]
[1]State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University
[2] Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing Jiaotong University
[3]Key Laboratory of Beijing for Railway Engineering, Beijing Jiaotong University
[4]Department of Computer Science, Durham University
[5] Frontiers Science Center for Smart High-speed Railway System, Beijing Jiaotong University
{21112046, hyin, lzyinxx}@bjtu.edu.cn, xingyu.miao@durham.ac.uk

Abstract

Accurate identification of non-rigid motion is crucial for geometric validity in self-supervised monocular depth estimation (MDE), yet it remains challenging for current methods. Inspired by Human Visual Perception (HVP), we present VFM-Dynamo, an efficient self-supervised MDE framework that disambiguates non-rigid objects by combining a vision foundation model (VFM) such as Grounding DINO and SAM with coarse to fine motion identification strategy: 1) Integrating visual-textual feature dependencies, we activate the capability of Grounding DINO global understanding with text prompts to detect potential object bounding boxes. 2) We then estimate coarse motion mask from reprojection error and photometric consistency to separate static from dynamic content. This mask provides coarse guidance for constraining rigid flow in static regions through two analogous estimates aligned by a consistency loss, and it drives a non-dynamical suppression module that removes boxes associated with static objects. 3) The remaining boxes are used to prompt SAM to produce refined motion identification, which in turn guide the joint optimization of depth and optical flow. Additionally, due to foreground–background mixing, standard interpolation-based upsampling often produces boundary artifacts. We introduce a learnable neighbor affinity interpolation (LNAI) module, which directly upsamples depth to full resolution and can be seamlessly integrated as a plug-and-play component. Experiments on a series of benchmarks demonstrate that proposed framework achieves state-of-the-art performance among unsupervised methods. Code is available at https://github.com/Qianqian3764/VFM-Dynamo.

1 Introduction

Monocular depth estimation (MDE) plays a fundamental role in computer vision, with broad applications in autonomous navigation, augmented reality, and video analysis. Previous supervised methods requires ground-truth depth annotations for training, whose collection is expensive and time-consuming. Consequently, self-supervised methods leverage synchronized stereo-pairs [Godard *et al.*, 2017] or geometric constraints [Zhou *et al.*, 2017] from monocular video as supervision signals, making monocular training more attractive.

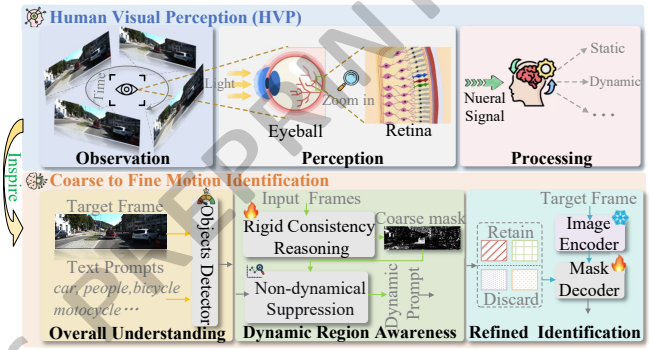

Figure 1: Our coarse to fine non-rigid motion identification pipeline inspired by human visual perception.

Current training pipeline of self-supervised MDE methods typically minimize geometry consistency loss to jointly estimate depth and camera pose [Godard *et al.*, 2019; Chang and Wang, 2025; Zheng *et al.*, 2025]. This pipeline relies on rigid geometric projection under static scene assumption. Despite overall high performance, the inconsistent geometric locations of moving objects lead to misaligned pixel matching, providing no longer valid supervision signals. To address this problem, a critical step is identifying dynamic objects from the rigid background. Several works leveraged semantic labels [Klingner *et al.*, 2020; Lee *et al.*, 2021; Dong *et al.*, 2024], optical flow [Yin and Shi, 2018; Liu *et al.*, 2019; Miao *et al.*, 2023; Zhou *et al.*, 2025], geometry consistency [Bian *et al.*, 2019; Godard *et al.*, 2019] or the discrepancy between single-frame and multi-frame depth [Watson *et al.*, 2021; Woo *et al.*, 2024] to discern dynamic objects. However, semantic labels involve needless static objects whereas geometry/discrepancy-induced methods fail to guarantee the accuracy in modeling dynamic objects. More fundamentally,

*Corresponding author.

motion identification is a complex and challenging problem that has been studied for decades, but there are still open questions [Goli *et al.*, 2025].

Recent research in object identification has been transformative, with the vision foundation model (VFM) such as Grounding DINO [Liu *et al.*, 2024b] and SAM [Kirillov *et al.*, 2023] emerging as a significant breakthrough. However, both of them face identification challenges if objects are momentarily motionless, and in distinguishing foreground objects from background 'noise'. This naturally raises the question:"Can we leverage the power of VFM for non-rigid moving object identification and advance self-supervised MDE ?"

As shown in Figure 1, we propose VFM-Dynamo, a progressive motion identification pipeline inspired by Human Visual Perception (HVP) for self-supervised MDE, which simulates HVP process and effectively tailors VFM from global perception to fine-grained attention. Specifically, as in early HVP stages, by computing cross-modal correlations between visual features and text embeddings, we first drive Grounding DINO to derive reliable object boxes, achieving global perception of potential moving objects. Second, we design rigid consistency reasoning to derive coarse motion clues, acting as a selector of moving regions within a frame. This strategy decouples rigid flow in two different manners and penalizes the inconsistency in the predicted rigid flow between them, indirectly providing sufficient regularization for dynamic objects. We then introduce non-dynamical suppression, which leverages coarse motion identification to filter out static object boxes. Finally, the remaining dynamic object boxes serve as prompts for SAM to refine motion identification, further optimizing the joint training of depth and optical flow network. In addition, aim to alleviates the edge-blurring problem in the traditional upsampling method, we leverage the spatial coherence in neighbor pixels and propose a learnable neighbor affinity interpolation (LNAI) module to estimate the interpolation pixels from a lower resolution directly. The main contributions of our work are summarized as follows:

- To the best of our knowledge, our framework is the first work that tailors VFM for accurately non-rigid motion identification in self-supervised MDE. Inspired by HVP, our approach introduces a coarse to fine learning framework that leverages the advantage of VFM and flow to mitigate motion bias in training process.
- We propose an upsampling strategy that is compatible with existing self-supervised MDE pipelines and achieves significant performance improvements.
- We evaluate VFM-Dynamo framework on a serious of widely used datasets and establish a new state-of-the-art performance in dynamic areas.

2 Related Work

Self-Supervised Depth Estimation. Self-supervised methods mitigate the reliance on expensive 3D sensors and compute geometric consistency between stereo images [Godard *et al.*, 2017] and monocular video frames [Zhou *et al.*, 2017] for training. Since Monodepth2 [Godard *et al.*, 2019] suggested a new benchmark with minimum reprojection loss, many advances adopt its methodology to improve depth predictions. MonoMixer [Chang and Wang, 2025] compensated for the limited receptive field of convolutions with Transformer, improving performance while maintaining a lightweight architecture. DiffSQL [Zheng *et al.*, 2025] leveraged the generative priors of stable diffusion for feature augmentation. Mono-ViFI [Liu *et al.*, 2024a] incorporated a flow-based video frame interpolation model for temporal augmentation.

Dynamic Scenes. To address the inaccurate training signals in dynamic scenes, identifying non-rigid motion from the rigid background is an essential step. MonoDepth2 [Godard *et al.*, 2019] introduced an auto-masking for moving objects. SC-Depth [Bian *et al.*, 2019] and SC-DepthV3 [Sun *et al.*, 2023] derived motion segmentation from the proposed self-discovered mask. SGDepth [Klingner *et al.*, 2020] and DynamicDepth [Feng *et al.*, 2022] employed precomputed segmentation masks to identify objects. DFNet [Zou *et al.*, 2018] and [Liu *et al.*, 2019] achieved this through the forward-backward consistency and rigidity reasoning of optical flow. Dynamo-Depth [Sun and Hariharan, 2023] jointly learned depth, optical flow, and motion segmentation to disambiguate dynamic motion. ManyDepth [Watson *et al.*, 2021] and ProDepth [Woo *et al.*, 2024] exploited the inconsistency between single-frame and multi-frame depth. While these methods are insightful, the reliance on expensive semantic labels often includes irrelevant static objects, and other approaches struggle to accurately estimate the motion of all dynamic objects in the real world.

VFM-based Depth Estimation. VFM, such as vision-language models (VLMs) [Radford *et al.*, 2021], Grounding DINO [Liu *et al.*, 2024b] and SAM [Kirillov *et al.*, 2023] are pre-trained on massive amounts of pairs of texts and images and has demonstrated powerful scene-reasoning abilities in computer vision tasks. DepthCLIP [Zhang *et al.*, 2022] is a pioneering attempt that leveraged the pretrained knowledge of CLIP for zero-shot depth estimation. Furthermore, subsequent studies [Hu *et al.*, 2024] and [Son and Lee, 2024] revealed that learnable prompts achieve better performance compared to manually designed templates. Hybriddepth [Zhang *et al.*, 2025b] enhanced feature representation with aggregated multi-grained features from CLIP and DINO under contrastive language guidance. They are still limited to ambiguity introduced by the moving objects. Our work extends the robust VFM framework to non-rigid motion identification for self-supervised MDE by adapting it with coarse to fine framework and dynamic prompts.

3 Method

We focus on solving for the depth estimation in the dynamic scene observed from a moving camera. Figure 2 presents an overview of our method. In the following, we will describe specific components in details.

3.1 Motivation and Background

Unsupervised Depth and Optical Flow. Unsupervised depth and flow estimation both rely on photometric consistency between the synthesized image $\hat{I}$ and the original image I, formulated as:

$$\mathcal{L}(\Theta) \sim \rho(\hat{I}(\Theta), I), \tag{1}$$

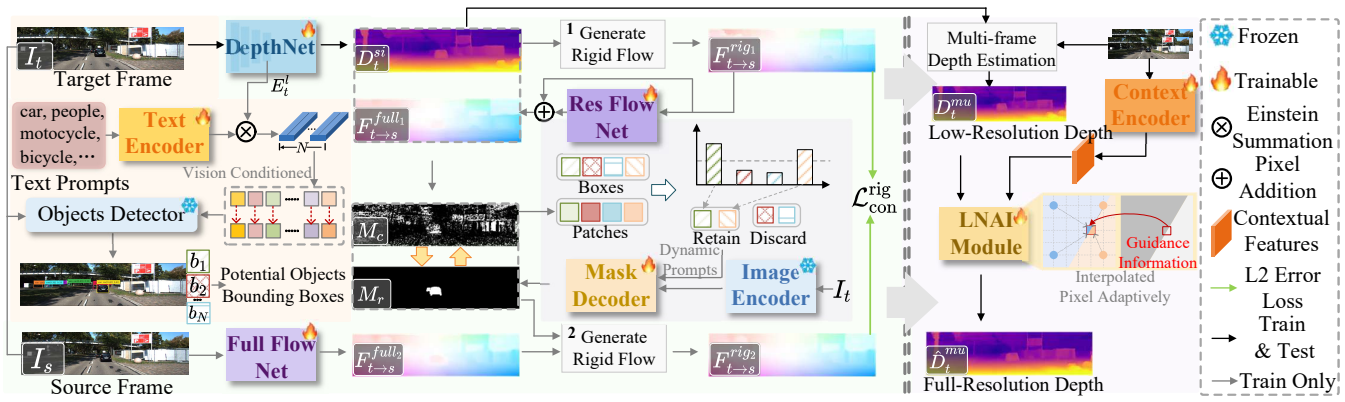

Figure 2: Overview of our VFM-Dynamo architecture. Given text prompts and target frame, we first generate object bounding boxes $\mathcal{B}_{obj}$ to achieve global scene understanding. Then, we calculate coarse mask M_c to provide guidance for two rigid flows aligned by the consistency loss $\mathcal{L}_{con}^{rig}$. Based on M_c and $\mathcal{B}_{obj}$, we design non-dynamical suppression to extract bounding boxes $\mathcal{B}_{dyn}$ that are most likely associated with non-rigid regions. These boxes are subsequently used to prompt SAM to produce a refined motion mask M_r, which updates M_c and facilitates self-supervised depth learning. Finally, the low resolution depth is upsampled using LNAI module to obtain full resolution depth.

where ρ is a measure of similarity calculated by $\mathcal{L}_1$ and SSIM, Θ is the model with learnable parameters.

Given a target image I_t and source image I_s, $s \in \{t-1, t+1\}$, depth and optical flow synthesize image in fundamentally different ways. In terms of depth, the image $\hat{I}$ can be synthesized by predicted depth D_t and camera pose $T_{t\to s}$:

$$\hat{I}_t^d(p) = I_s\langle \text{proj}(D_t, T_{t\to s}, K)\rangle, \tag{2}$$

where p is the pixel coordinates and K is camera intrinsics. In terms of optical flow, the image $\hat{I}$ can be synthesized by predicted flow field $F_{t\to s}$:

$$\hat{I}_t^{opt}(p) = I_s\langle p + F_{t\to s}(p)\rangle. \tag{3}$$

In this case, optical flow reconstructs images via per-pixel motion warping, which naturally accommodates dynamic objects, whereas self-supervised depth estimation relies on rigid geometric projection/epipolar constraints, making it sensitive to violations caused by dynamic objects. Therefore, as optical flow is adept at tracking the movement of objects, this advantage motivates us to explore *how such motion cues can be exploited during self-supervised depth training to provide effective regularization for dynamic regions.*

Human Visual Perception. In visual cognition, humans tend to follow a global-to-local perception paradigm, in which a visual image is first perceived as an integrated whole and subsequently analyzed in terms of its constituent parts [Xia *et al.*, 2024]. This observation motivates us to investigate *whether we can mimic the staged process of HVP: global scene understanding to motion clues-driven region localization and refined non-rigid motion identification.*

3.2 Coarse to Fine Motion Identification

To properly exploit the advantages of optical flow for dynamic objects, we reason rigid flow of static scene in two different manners and enforce consistency constraints between them. This process depends on accurate identification of moving objects, which is hindered by the complexity and time-varying nature of object motion. Inspired by HVP, we introduce VFM and design a coarse to fine motion identification strategy to effectively address this issue.

Overall Understanding. As in early HVP stages, we harness open-vocabulary 2D detector–Grounding DINO [Liu *et al.*, 2024b] to access the overall scene understanding. Relevant text prompts, such as "car, people, motorcycle, bicycle,...," are defined and fed into trainable text encoder to output text embeddings $T = [T_1, ..., T_N] \in \mathbb{R}^{N\times C}$. To ensure that the text embeddings are more similar to its corresponding visual features, text-visual similarity is computed as:

$$S_t = E_t^l \otimes T \in \mathbb{R}^{N\times H\times W}, \tag{4}$$

where $E_t^l \in \mathbb{R}^{C\times H\times W}$ is visual features from the last layer of DepthNet encoder. $\otimes$ is einstein summation over channel dimension. We normalize it along spatial dimension and aggregate visual features to derive text-guided visual features:

$$E_t^{'} = S_t \otimes E_t^l \in \mathbb{R}^{N\times C}, \tag{5}$$

where $E_t^{'}$ represents the aggregated visual features corresponding to each text token.

Finally, we adaptively update the text embeddings via a gated residual fusion:

$$\hat{T} = T + \sigma\left(\frac{T\cdot E_t^{'}}{\sqrt{C}}\right)\cdot E_t^{'} \in \mathbb{R}^{N\times C}, \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Then the vision conditioned text embeddings $\hat{T}$ are adopted to predict bounding boxes $\mathcal{B}_{obj} = \{b_i\}_{i=1}^N$ for all referred objects.

Non-rigid Motion Awareness. We input I_t into the depth network $\mathcal{D}$ to estimate the depth D_t^{si} and concat I_t and I_s as input to the PoseNet $\mathcal{P}$, from which we estimate the relative camera pose $T_{t\to s}$. Then the rigid flow that purely induced by the camera motion can be obtained:

$$\hat{p}_s = KT_{t\to s}D_t^{si}(p_t)\,K^{-1}p_t, \quad F_{t\to s}^{rig_1}(p_t) = \hat{p}_s - p_t, \tag{7}$$

where $F_{t\to s}^{rig_1}$ describes the motion of a rigid scene projected onto image plane. Afterwards, we cascade ResFlowNet $\mathcal{R}$ to learn non-rigid residual that capture object motion:

$$F_{t\to s}^{non} = \mathcal{R}(I_t, I_s, \hat{I}_t^{opt_r}, F_{t\to s}^{rig_1}, \rho(I_t, \hat{I}_t^{opt_r})), \tag{8}$$

where $\hat{I}_t^{opt_r}$ denotes the synthesize image calculated by $F_{t\to s}^{rig_1}$, The final full flow prediction $F_{t\to s}^{full_1} = F_{t\to s}^{rig_1} + F_{t\to s}^{non}$. According to Equation 2 and 3, the synthetic image $\hat{I}_t^d$ and $\hat{I}_t^{opt}$ are reconstructed using D_t^{si} and $F_{t\to s}^{full_1}$ respectively, and then calculate the difference with the target image:

$$\rho(\hat{I}, I) = \alpha \frac{1 - \text{SSIM}(\hat{I}, I)}{2} + (1-\alpha)\|\hat{I} - I\|_1, \tag{9}$$

where α is set to 0.85. Instead of averaging differences over all source images, we use the minimum and predict coarse motion mask:

$$M_c = \left[\min_s \rho\left(I_t, \hat{I}_t^{opt}\right) < \min_s \rho\left(I_t, \hat{I}_t^{d}\right)\right], \tag{10}$$

where $[\,]$ is the Iverson bracket. In cases where binary mask M_c is 1 in regions considered to be dynamic, and 0 otherwise.

And then we predict the full flow $F_{t\to s}^{full_2}$ via concating and passing I_t and I_s to a FullflowNet $\mathcal{F}$. The rigid flow $F_{t\to s}^{rig_2}$ is computed in a manner gated by the motion mask M_c:

$$F_{t\to s}^{rig_2} = F_{t\to s}^{full_2}(1 - M_c). \tag{11}$$

The consistency between the rigid optical flows generated from two distinct manners is enforced:

$$\mathcal{L}_{con}^{rig} = \left\| F_{t\to s}^{rig_1} - f_{t\to s}^{rig_2} \right\|_2, \tag{12}$$

where $\|\ \|_2$ represents $\mathcal{L}_2$ norm.

In real-world scenes, the prevalence of low-texture regions and frequent occlusions makes it difficult to strictly satisfy the aforementioned criteria in Equation 10. As shown in the third row of Figure 3, M_c inevitably contains substantial noise, preventing consistency constraint in Equation 12. SAM [Kirillov *et al.*, 2023] is exploited to refine coarse motion cues and obtain accurate motion masks.

Refined Identification. With coarse motion mask M_c, we measure overlap ratio and suppress non-dynamical objects:

$$\mathcal{B}_{dyn} = \left\{ b_i \in \mathcal{B}_{obj} \middle| \frac{\sum_{p\in\Omega(b_i)} M_c(p)}{|\Omega(b_i)|} \geq \tau \right\}, \tag{13}$$

where $|\Omega(b_i)|$ is the area of box b_i, τ is the threshold and we set 0.45. Each candidate box b_i, specified by its four corner coordinates, is evaluated by computing its overlap ratio over the enclosed patches in M_c.

Finally, we harness the dynamic bounding box to prompt SAM with a trainable mask decoder:

$$M_r = SAM(I_t, \mathcal{B}_{dyn}), \tag{14}$$

where M_r is the refined motion identification, which substitutes M_c to facilitate rigid consistency reasoning on more reliable static regions.

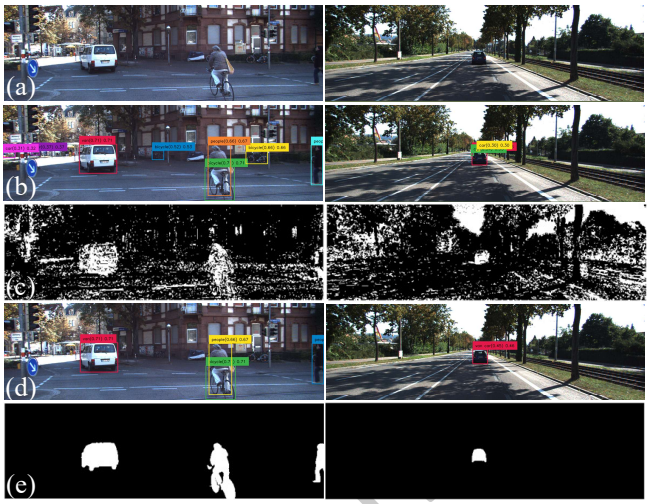

Figure 3: Illustration examples of our coarse to fine non-rigid motion identification. (a) Target frame, (b) potential bounding boxes $\mathcal{B}_{obj}$ of overall understanding. (c) Coarse motion mask M_c of rigid consistency reasoning. (d) Then it serves as moving regions selector to generate dynamic prompts $\mathcal{B}_{dyn}$, guiding SAM to identify and localize the moving objects M_r (e).

3.3 Learnable Neighbor Affinity Interpolation

Most existing methods upsample low-resolution predictions to full resolution using fixed interpolation schemes (e.g., nearest-neighbor and bilinear) which often result in boundary artifacts. To alleviate this problem, we propose a Learnable Neighbor Affinity Interpolation (LNAI) module.

After general multi-frame depth estimation, the regressed depth $D_t^{mu} \in \mathbb{R}^{H\times W\times 1}$ could be obtained. We extracted contextual features $\Psi \in \mathbb{R}^{H\times W\times C}$ from target frame, where $C = 144$. Our interpolation step can be defined as:

$$\hat{D}_t^{mu} = \mathcal{U}(D_t^{mu}, \Psi), \mathcal{U} : \mathbb{R}^{H\times W\times 1} \to \mathbb{R}^{4H\times 4W\times 1}, \tag{15}$$

where $\mathcal{U}$ denotes LANI module. $\hat{D}_t^{mu}$ is interpolated full resolution depth and D_t^{mu} is multi-frame depth at $\frac{1}{4}$-resolution.

As shown in Figure 4, we first enriched Ψ with RoPE [Su *et al.*, 2024] positional embeddings and derive matrices Y_1 and Y_2 through projection layer:

$$Y_1 = W_1 \cdot \Psi, Y_2 = W_2 \cdot \Psi. \tag{16}$$

For $Y_1 \in \mathbb{R}^{H\times W\times 144}$, we divide the channels into 16 groups, with 9 channels in each group and reshape it to the size of $HW \times 16 \times 1 \times 9$. Then for $Y_2 \in \mathbb{R}^{H\times W\times 144}$, we similarly group the features along the channel dimension and reshape it into size of $16\times 9\times H\times W$. Afterwards, we divide Y_2 into $HW \times 16 \times 9$ local windows with the size of $N \times N$, and reshape it to the size of $HW \times 16 \times 9 \times N^2$, the neighbor affinity $\mathcal{A}$ can be computed as:

$$\mathcal{A}_{(k,g,n)} = \sum_{l=1}^{C} Y_1(x_{k,g,l}) \cdot Y_2^T(x_{g,n,l}), \tag{17}$$

where $k \in \{1, 2, \ldots, HW\}$ is the index of contextual features, $g \in \{1, 2, \ldots, 16\}$ is the index of feature group. $n \in$

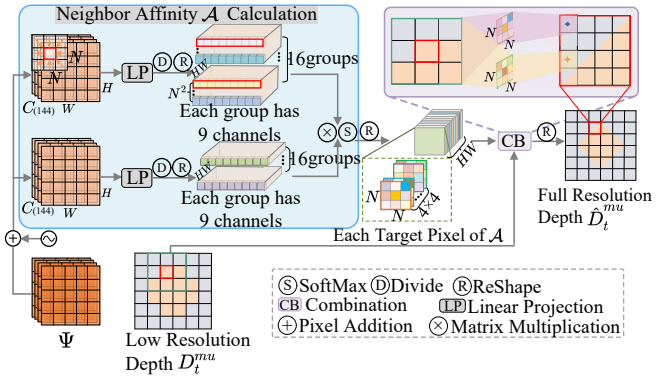

Figure 4: The implementation details of proposed LNAI module.

$\{1, 2, \ldots, N^2\}$ is the index of neighbors, $l \in \{1, 2, \ldots, 9\}$ is the index of channel within each group. $x_{k,g,l}$ and $x_{g,n,l}$ are the points in the local window $\Omega \in N^2$. Then we execute softmax on $\mathcal{A}$ and reshape it into size of $N^2 \times 4 \times 4 \times H \times W$.

To align with the $N \times N$ affinity $\mathcal{A}$, we divide D_t^{mu} into HW local windows with the size of $N \times N$ and reshape it to the size of $N^2 \times 1 \times 1 \times H \times W$. After that, we yield full resolution depth:

$$\hat{D}_t^{mu} = \mathcal{A} \oplus D_t^{mu} \in \mathbb{R}^{1\times 4\times 4\times H\times W}, \tag{18}$$

where $\oplus$ is combination operation, each target interpolated pixel is weighted by a window that is not shared by another pixel. Finally, we reshape the full resolution depth to the size of $4H \times 4W \times 1$.

3.4 Loss Functions

Flow Photometric Loss. We adopt the same measurement in Equation 9 to compute the photometric loss of full flow $F_{t\to s}^{full_1}$ and $F_{t\to s}^{full_2}$:

$$\mathcal{L}_p^{opt} = \min_s \rho\left(\hat{I}_t^{opt}, I_t\right). \tag{19}$$

Depth Loss. For single-frame depth D_t^{si} and multi-frame depth $\hat{D}_t^{mu}$, we similarly compute photometric loss and use smoothness loss [Godard *et al.*, 2017] $\mathcal{L}_{sm}$ to regularize:

$$\mathcal{L}_p^{d} = \mu \cdot \min_s \rho\left(\hat{I}_t^{d}, I_t\right) + \gamma\mathcal{L}_{sm}, \tag{20}$$

where μ is auto-mask [Godard *et al.*, 2019] and $\gamma = 0.001$.

Final Loss. In summary, the final loss function becomes a weighted sum of the above terms:

$$\mathcal{L}_{total} = \mathcal{L}_p^{d_{si}} + \mathcal{L}_p^{d_{mu}} + \lambda_1\mathcal{L}_{con}^{rig} + \lambda_2\mathcal{L}_p^{opt_1} + \lambda_3\mathcal{L}_p^{opt_2}, \tag{21}$$

where $\mathcal{L}_p^{d_{si}}$ and $\mathcal{L}_p^{d_{mu}}$ is the loss of D_t^{si} and $\hat{D}_t^{mu}$. λ_1, λ_2 and λ_3 are the loss weights and we set 0.001, 0.1, 0.1 respectively.

4 Experiments

We evaluate our method on three challenging datasets: KITTI [Geiger *et al.*, 2013], DDAD [Guizilini *et al.*, 2020] and Cityscapes [Cordts *et al.*, 2016]. We use standard metrics to measure model performance. For dynamic regions, we follow the protocol from SC-DepthV3 [Sun *et al.*, 2023].

4.1 Datasets

KITTI. We adopt Eigen split [Eigen *et al.*, 2014] of KITTI, comprising 39,810 training images and 4,424 validation images. The test set includes 697 raw LiDAR frames and 652 sparsity-corrected ground truth images [Uhrig *et al.*, 2017]. Depth is capped at 80m and images are resized to 640×192.

DDAD. DDAD is a challenging driving dataset with more dynamic objects. It contains 3,950 test frames, with depth capped at 200m and image resolution set to 640×384.

Cityscapes. Cityscapes is a benchmark designed for urban scene understanding, including a substantial number of dynamic objects and containing 1,525 test images. The depth range is capped at 80m and images are resized to 512×192.

4.2 Implementation Details

We use two NVIDIA RTX 3090 GPUs and Pytorch with batch size 6 for all training experiments. We employ the AdamW optimizer to minimize the loss functions and train the network for 20 epochs. For our baseline model, the DepthNet $\mathcal{D}$ follows the same architecture in [Zhao *et al.*, 2022]. The remaining network components adopt the design of [Watson *et al.*, 2021]. We unfreeze PWCNet [Sun *et al.*, 2018] as FullflowNet $\mathcal{F}$ in our framework. ResFlowNet $\mathcal{R}$ follows the same architecture in [Miao *et al.*, 2023] and a pre-trained ResNet18 is used as our context encoder. Other experimental settings are similar with [Watson *et al.*, 2021].

4.3 Results on KITTI

As shown in Table 1, for raw data, compared with DiffSQL [Zheng *et al.*, 2025], our Abs Rel and RMSE error improve 8.5% and 4.1%. Compared with motion focused method DualDepth [Feng *et al.*, 2025], our Abs Rel error improves 5.5%. When compared with Mono-ViFI [Liu *et al.*, 2024a], our method performs best on the Abs Rel, RMSE_{log}, δ_1, and δ_2. Since Mono-ViFI leverages a pretrained flow-based video frame interpolation model during inference to capture temporal information, it requires access to future frames that are not yet observed when making current predictions. For dynamic areas, existing motion focused methods fail to sustain consistent performance and show performance drop while our method achieves state-of-the-art results. Moreover, since raw LiDAR data is not well suited for reliably evaluating performance in dynamic areas, we further assess our method using the sparsity-corrected ground truth [Uhrig *et al.*, 2017]. VFM-Dynamo outperforms other comparison methods on almost metrics for both full image and dynamic areas.

Qualitative results are shown in Figure 5, DS-Depth and Dynamo are two motion focused methods that utilize optical flow to introduce motion priors. The last two cases involve dynamic objects (yellow boxes), while comparison methods fail to estimate accurate depth. In the first case, our methods preserves noticeably sharper depth around object boundaries.

Figure 6 shows visual results for the two types of motion identification. MonoDepth2 [Godard *et al.*, 2019] and ManyDepth [Watson *et al.*, 2021] generate binary mask sometimes fails to (i) identify moving objects (e.g., missing car and cyclist in first two cases). ProDepth [Woo *et al.*, 2024] produces soft mask but cannot (ii) distinguish between multiple objects (e.g., entangled car segmentation in the last case), while our

	Method	Venue	Train	Full Image							Dynamic		
				Abs Rel ↓	Sq Rel ↓	RMSE ↓	$RMSE_{log}$ ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	Abs Rel ↓	RMSE ↓	δ_1 ↑
Raw GT													
NMF	Lite-Mono [Zhang *et al.*, 2023]	CVPR'23	M	0.107	0.765	4.561	0.183	0.886	0.963	0.983	0.185	7.301	0.760
NMF	SC-DepthV3 [Sun *et al.*, 2023]	TPAMI'23	M	0.118	0.756	4.709	0.188	0.864	0.960	0.984	0.205	-	0.703
NMF	MonoDiffusion [Shao *et al.*, 2024]	TCSVT'24	M	0.103	0.720	4.412	0.178	0.894	0.965	0.984	0.184	7.218	0.760
NMF	GeoDepth [Wu *et al.*, 2025]	CVPR'25	M	0.100	0.694	4.381	0.176	0.897	0.966	0.984	-	-	-
NMF	Mono-ViFI [Liu *et al.*, 2024a]	ECCV'24	M	0.091	**0.589**	**4.088**	0.166	0.912	0.969	**0.986**	0.188	7.247	0.746
NMF	DCPI-Depth [Zhang *et al.*, 2025a]	TIP'25	M	0.095	0.662	4.274	0.170	0.902	0.967	0.985	-	-	-
NMF	Hybrid-depth [Zhang *et al.*, 2025b]	ICCV'25	M	0.093	0.596	4.113	0.167	0.910	0.970	**0.986**	-	-	-
NMF	DiffSQL [Zheng *et al.*, 2025]	IJCAI'25	MS	0.094	0.684	4.306	0.169	0.902	0.967	0.984	-	-	-
MF	Dynamo [Sun and Hariharan, 2023]	NIPS'23	M	0.112	0.758	4.505	0.183	0.873	0.959	0.984	0.211	7.314	0.713
MF	DS-Depth [Miao *et al.*, 2023]	TCSVT'23	M	0.095	0.698	4.329	0.173	0.905	0.966	0.984	0.192	7.457	0.742
MF	Manydepth2 [Zhou *et al.*, 2025]	RAL'25	M	0.091	0.649	4.232	0.170	0.909	0.968	0.984	0.169	6.880	0.778
MF	DualDepth [Feng *et al.*, 2025]	TIP'25	M	0.091	0.703	4.215	0.171	0.911	0.967	0.984	-	-	-
MF	VFM-Dynamo (Ours)	This paper	M	**0.086**	0.642	4.129	**0.165**	**0.916**	**0.971**	**0.986**	**0.165**	**6.740**	**0.780**
Improved GT													
NMF	MonoDiffusion [Shao *et al.*, 2024]	TCSVT'24	M	0.073	0.377	3.451	0.115	0.935	0.988	0.997	0.122	4.955	0.853
NMF	AQUANet [Bello *et al.*, 2024]	TIP'24	M	0.073	0.374	3.572	0.115	0.935	0.985	0.996	-	-	-
NMF	Mono-ViFI [Liu *et al.*, 2024a]	ECCV'24	M	0.066	**0.301**	3.120	0.103	0.951	0.992	**0.998**	0.125	4.879	0.841
NMF	DiffSQL [Zheng *et al.*, 2025]	IJCAI'25	MS	0.065	0.338	3.259	0.105	0.947	0.990	**0.998**	-	-	-
MF	Dynamo [Sun and Hariharan, 2024]	NIPS'23	M	0.087	0.454	3.685	0.130	0.917	0.985	0.997	0.147	5.261	0.808
MF	DS-Depth [Miao *et al.*, 2023]	TSCVT'23	M	0.067	0.358	3.303	0.108	0.946	0.989	0.997	0.130	5.252	0.831
MF	Manydepth2 [Zhou *et al.*, 2025]	RAL'25	M	0.070	0.399	3.456	0.113	0.941	0.989	0.997	0.147	4.842	0.817
MF	VFM-Dynamo (Ours)	This paper	M	**0.057**	0.329	**3.116**	**0.098**	**0.956**	**0.993**	**0.998**	**0.110**	**4.592**	**0.871**

Table 1: Results on KITTI [Geiger *et al.*, 2013] using the raw and improved ground truth. NMF indicates that the method does not focus on motion identification, MF indicates that the method focus on motion identification. M denotes monocular training, MS combines monocular videos with stereo pairs. The best results are in **bold**, and second best are underlined.

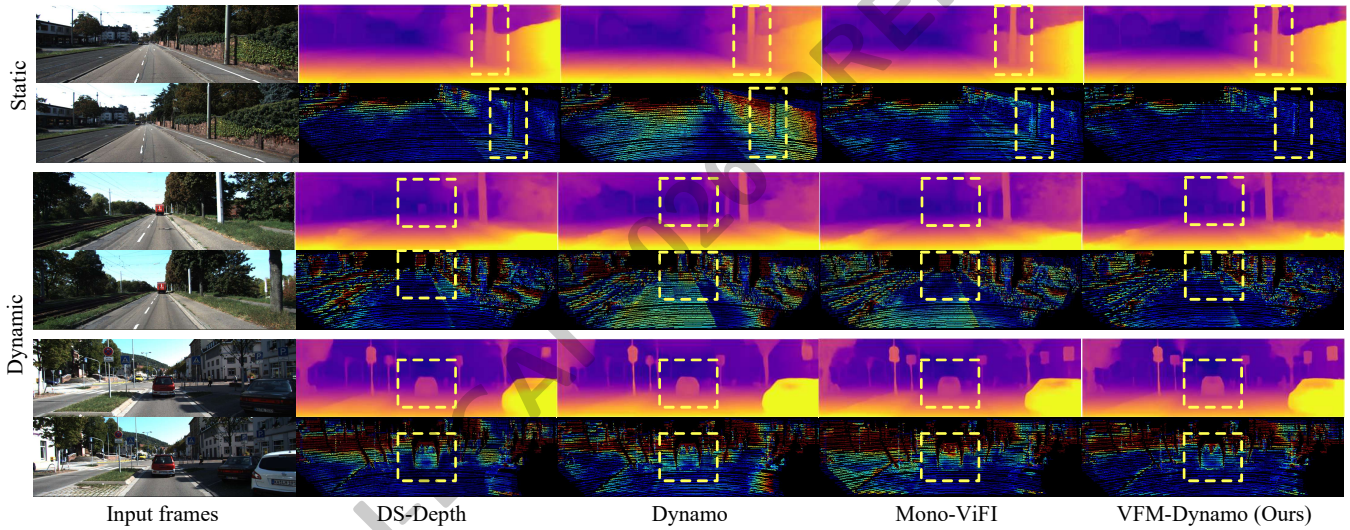

Figure 5: Qualitative results in cases of dynamic targets and object boundary. From top to bottom: the depth and reconstruction error maps. Input frames includes target and source frame. Error maps show the Abs Rel error compared to GT, the closer to the red, the greater the error.

method incorporates motion prior with dynamic prompts, resulting in the accurate identification of moving objects.

4.4 Results on DDAD and Cityscapes

The DDAD and Cityscapes datasets include a significant number of moving objects than KITTI because almost vehicles and pedestrian are moving on the road. As depicted in Table 2, VFM-Dynamo achieve superior performance compared with other methods, which demonstrates the excellent zero-shot generalization ability of our model.

4.5 Ablation Study

Our ablation study includes coarse to fine motion identification and LNAI module. All variants are trained on KITTI. The proposed components and their variants are integrated into a unified baseline model for fair comparison.

Effectiveness of coarse to fine. In Table 3, full flow $F_{t\to s}^{full_1}$ calculated from depth improves performance as it introduces additional geometric constraints (row 2). We then add $\mathcal{L}_{con}^{rig}$ calculated from coarse mask (row 3), providing attention for dynamic areas indirectly. Furthermore, we adapt VFM for fine-grained motion mask. Removing dynamic prompts (DP) from coarse mask leads to performance drop (row 4) as the mask output by SAM inevitably contains static objects that

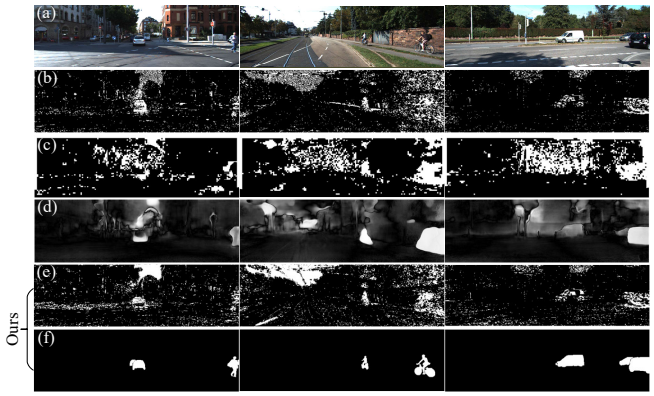

Figure 6: Comparison results of non-rigid motion identification. (a) Target frame, (b) MonoDepth2, (c) ManyDepth, (d) ProDepth, (e) and (f) are our coarse and refined motion mask.

Data	Methods	Full Image			Dynamic	
		Abs Rel ↓	RMSE ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑
DDAD	MonoDiffusion	0.254	19.916	0.544	0.246	0.531
	DS-Depth	0.297	19.115	0.521	0.264	0.525
	ProDepth	0.263	19.432	0.524	0.278	0.451
	ManyDepth2	0.236	18.270	0.348	0.262	0.505
	VFM-Dynamo	**0.210**	**18.116**	**0.623**	**0.220**	**0.586**
Cityscapes	MonoDiffusion	0.152	8.111	0.796	0.112	0.875
	DS-Depth	0.170	8.461	0.741	0.166	0.744
	ProDepth	0.146	**7.706**	0.801	0.143	0.813
	ManyDepth2	0.157	8.001	0.774	0.124	0.851
	VFM-Dynamo	**0.140**	8.015	**0.808**	**0.101**	**0.894**

Table 2: Results on DDAD [Guizilini *et al.*, 2020] and Cityscapes [Cordts *et al.*, 2016] datasets.

Stage	Setting	Full Image			Dynamic	
		Abs Rel ↓	RMSE ↓	δ_1 ↑	Abs Rel ↓	δ_1 ↑
Baseline		0.092	4.479	0.910	0.252	0.722
Coarse (M_c)	w/o $\mathcal{L}_{con}^{rig}$	0.090	4.415	0.911	0.223	0.731
	w/ $\mathcal{L}_{con}^{rig}$	0.090	4.404	0.912	0.217	0.736
Fine (M_r)	w/o DP	0.091	4.453	0.909	0.236	0.730
	w/o VC	0.090	4.301	0.913	0.197	0.752
	Full	**0.088**	**4.203**	**0.915**	**0.184**	**0.760**

Table 3: Results of coarse to fine motion identification.

hinder the rigid consistency constraint. Vision conditioned (VC) textual features integrate scene-aware visual features, yielding better performance than text features alone (row 5).

Necessity of accurate non-rigid motion identification. As shown in Table 4, we apply the proposed method to refine the non-rigid motion identification of baseline approaches. For ProDepth [Woo *et al.*, 2024], which uses a soft motion mask, we introduce a thresholding operation (threshold = 0.5) to adapt it to our framework. All methods show significant performance improvements, demonstrating that accurate motion identification effectively mitigates interference from erroneous geometric correspondences during training.

Universality of LNAI module. We adopt single-frame methods, including MonoDepth2 [Godard *et al.*, 2019] and Lite-mono [Zhang *et al.*, 2023], the multi-frame method ManyDepth [Watson *et al.*, 2021] as baselines. For MonoDepth2 and ManyDepth, we remove the decoder of DepthNet from the last two scales, and for Lite-mono, we remove it from the last scale, retaining depth outputs only at $\frac{1}{8}$, $\frac{1}{4}$, and $\frac{1}{4}$ resolution, respectively. Subsequently, we integrate the LNAI module to upsample the $\frac{1}{4}$-resolution depth maps directly to full resolution. As shown in Table 5, the LNAI module consistently improves nearly all metrics across all baseline methods. Moreover, by eliminating the final decoding stages, the introduction of LNAI incurs almost no additional inference overhead, highlighting its potential to enhance the efficiency of approaches that employ complex decoders.

Methods	Dynamic			
	Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1 ↑
MonoDepth2	0.204	3.096	8.052	0.735
+RM	0.197 (↑3.43%)	3.010 (↑2.78%)	7.741 (↑3.86%)	0.740 (↑0.68%)
ManyDepth	0.206	3.170	7.908	0.734
+RM	0.202 (↑1.94%)	2.907 (↑8.30%)	7.552 (↑4.50%)	0.736 (↑0.27%)
ProDepth	0.182	2.820	7.455	0.766
+RM	0.178 (↑2.20%)	2.772 (↑1.70%)	7.375 (↑1.07%)	0.766 (↑0.00%)

Table 4: Results of different methods combined with refined motion identification (RM).

Methods	Time (s)	Full Image			
		Abs Rel ↓	Sq Rel ↓	RMSE ↓	δ_1 ↑
MonoDepth2	0.0301	0.115	0.903	4.863	0.877
+LNAI	+0.008	0.113 (↑1.74%)	0.849 (↑5.98%)	4.854 (↑0.19%)	0.879 (↑0.23%)
ManyDepth	0.0433	0.098	0.770	4.459	0.900
+LNAI	+0.006	0.097 (↑1.02%)	0.699 (↑9.22%)	4.431 (↑0.63%)	0.901 (↑0.11%)
Lite-Mono	0.0313	0.107	0.765	4.561	0.886
+LNAI	+0.002	0.106 (↑0.93%)	0.760 (↑0.65%)	4.540 (↑0.46%)	0.889 (↑0.34%)

Table 5: Results of different methods with the LNAI module.

Metric	DS-Depth	ProDepth	ManyDepth2	DualDepth	Ours
Params (M)	73.03	39.80	39.79	53.30	47.59
Time (s)	0.1103	0.1080	0.1802	-	0.1151

Table 6: Complexity comparison results. The batch size is 1.

Model complexity. In Table 6, our proposed model attains a good balance between model size and runtime speed. Notice that VFM-Dynamo outperforms the recent well-designed models ManyDepth2 [Zhou *et al.*, 2025] and DualDepth [Feng *et al.*, 2025] both in runtime and accuracy (Table 1).

4.6 Conclusion

In this paper, we present VFM-Dynamo, a novel self-supervised MDE methods inspired by HVP and designed a coarse to fine motion identification strategy. Additionally, we facilitate the direct upsampling of low-resolution depth to full-resolution using neighbor affinity and proposed approach can be seamlessly integrated into existing self-supervised methods. In future work, we would like to utilize VFM to distinguish the dynamic objects in the end-to-end manner.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant (2025YFF0519403) and National Natural Science Foundation of China (U2468223, 51827813).

References

[Bello *et al.*, 2024] Juan Luis Gonzalez Bello, Jaeho Moon, and Munchurl Kim. Self-supervised monocular depth estimation with positional shift depth variance and adaptive disparity quantization. *IEEE Transactions on Image Processing*, 33:2074–2089, 2024.

[Bian *et al.*, 2019] Jiawang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32, 2019.

[Chang and Wang, 2025] Zhiyong Chang and Yan Wang. Monomixer: Marrying convolution and vision transformer for efficient self-supervised monocular depth estimation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 756–764, 2025.

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

[Dong *et al.*, 2024] Yue-Jiang Dong, Fang-Lue Zhang, and Song-Hai Zhang. Mal: motion-aware loss with temporal and distillation hints for self-supervised depth estimation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7318–7324. IEEE, 2024.

[Eigen *et al.*, 2014] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[Feng *et al.*, 2022] Ziyue Feng, Liang Yang, Longlong Jing, Haiyan Wang, YingLi Tian, and Bing Li. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In *European Conference on Computer Vision*, pages 228–244. Springer, 2022.

[Feng *et al.*, 2025] Cheng Feng, Congxuan Zhang, Zhen Chen, Weiming Hu, Ke Lu, and Liyue Ge. Self-supervised monocular depth estimation with dual-path encoders and offset field interpolation. *IEEE Transactions on Image Processing*, 2025.

[Geiger *et al.*, 2013] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.*, 32(11):1231–1237, 2013.

[Godard *et al.*, 2017] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 270–279, 2017.

[Godard *et al.*, 2019] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proc. IEEE Int. Conf. Comput. Vis*, pages 3828–3838, 2019.

[Goli *et al.*, 2025] Lily Goli, Sara Sabour, Mark Matthews, Marcus A Brubaker, Dmitry Lagun, Alec Jacobson, David J Fleet, Saurabh Saxena, and Andrea Tagliasacchi. Romo: Robust motion segmentation improves structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6155–6164, 2025.

[Guizilini *et al.*, 2020] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020.

[Hu *et al.*, 2024] Xueting Hu, Ce Zhang, Yi Zhang, Bowen Hai, Ke Yu, and Zhihai He. Learning to adapt clip for few-shot monocular depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5594–5603, 2024.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[Klingner *et al.*, 2020] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 582–600. Springer, 2020.

[Lee *et al.*, 2021] Seokju Lee, Sunghoon Im, Stephen Lin, and In So Kweon. Learning monocular depth in dynamic scenes via instance-aware projection consistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1863–1872, 2021.

[Liu *et al.*, 2019] Liang Liu, Guangyao Zhai, Wenlong Ye, and Yong Liu. Unsupervised learning of scene flow estimation fusing with local rigidity. In *IJCAI*, pages 876–882, 2019.

[Liu *et al.*, 2024a] Jinfeng Liu, Lingtong Kong, Bo Li, Zerong Wang, Hong Gu, and Jinwei Chen. Mono-vifi: A unified learning framework for self-supervised single and multi-frame monocular depth estimation. In *Proc. Eur. Conf. Comput. Vis.*, pages 90–107, 2024.

[Liu *et al.*, 2024b] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

[Miao *et al.*, 2023] Xingyu Miao, Yang Bai, Haoran Duan, Yawen Huang, Fan Wan, Xinxing Xu, Yang Long, and Yefeng Zheng. Ds-depth: Dynamic and static depth estimation via a fusion cost volume. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Shao *et al.*, 2024] Shuwei Shao, Zhongcai Pei, Weihai Chen, Dingchi Sun, and Peter CY Chen. Monodiffusion: self-supervised monocular depth estimation using diffusion model. *IEEE Trans. Circuits Syst. Video Technol.*, 35(4):3664–3678, 2024.

[Son and Lee, 2024] Eunjin Son and Sang Jun Lee. Cabins: Clip-based adaptive bins for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4557–4567, 2024.

[Su *et al.*, 2024] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[Sun and Hariharan, 2023] Yihong Sun and Bharath Hariharan. Dynamo-depth: Fixing unsupervised depth estimation for dynamical scenes. *Advances in Neural Information Processing Systems*, 36:54987–55005, 2023.

[Sun and Hariharan, 2024] Yihong Sun and Bharath Hariharan. Dynamo-depth: fixing unsupervised depth estimation for dynamical scenes. *Advances in Neural Information Processing Systems*, 36, 2024.

[Sun *et al.*, 2018] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 8934–8943, 2018.

[Sun *et al.*, 2023] Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Uhrig *et al.*, 2017] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017.

[Watson *et al.*, 2021] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1164–1174, 2021.

[Woo *et al.*, 2024] Sungmin Woo, Wonjoon Lee, Woo Jin Kim, Dogyoon Lee, and Sangyoun Lee. Prodepth: Boosting self-supervised multi-frame monocular depth with probabilistic fusion. In *Proc. Eur. Conf. Comput. Vis.*, pages 201–217, 2024.

[Wu *et al.*, 2025] Haifeng Wu, Shuhang Gu, Lixin Duan, and Wen Li. Geodepth: From point-to-depth to plane-to-depth modeling for self-supervised monocular depth estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11525–11535, 2025.

[Xia *et al.*, 2024] Weihao Xia, Raoul De Charette, Cengiz Oztireli, and Jing-Hao Xue. Dream: Visual decoding from reversing human visual system. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8226–8235, 2024.

[Yin and Shi, 2018] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018.

[Zhang *et al.*, 2022] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6868–6874, 2022.

[Zhang *et al.*, 2023] Ning Zhang, Francesco Nex, George Vosselman, and Norman 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*, pages 18537–18546, 2023.

[Zhang *et al.*, 2025a] Mengtan Zhang, Yi Feng, Qijun Chen, and Rui Fan. Dcpi-depth: Explicitly infusing dense correspondence prior to unsupervised monocular depth estimation. *IEEE Transactions on Image Processing*, 2025.

[Zhang *et al.*, 2025b] Wenyao Zhang, Hongsi Liu, Bohan Li, Jiawei He, Zekun Qi, Yunnan Wang, Shengyang Zhao, Xinqiang Yu, Wenjun Zeng, and Xin Jin. Hybrid-grained feature aggregation with coarse-to-fine language guidance for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6678–6692, 2025.

[Zhao *et al.*, 2022] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *Proc. IEEE Int. Conf. 3D Vis.*, pages 668–678. IEEE, 2022.

[Zheng *et al.*, 2025] Heyuan Zheng, Yunji Liang, Lei Liu, and Zhiwen Yu. Diffsql: leveraging diffusion model for zero-shot self-supervised monocular depth estimation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 8823–8831, 2025.

[Zhou *et al.*, 2017] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1851–1858, 2017.

[Zhou *et al.*, 2025] Kaichen Zhou, Jia-Wang Bian, Jian-Qing Zheng, Jiaxing Zhong, Qian Xie, Niki Trigoni, and Andrew Markham. Manydepth2: Motion-aware self-supervised monocular depth estimation in dynamic scenes. *IEEE Rob. Autom. Lett.*, pages 1–8, 2025.

[Zou *et al.*, 2018] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 36–53, 2018.