Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GeoDiffusion:Enhancing Self-Supervised Monocular Depth Estimation in Dynamic Scenes

Fang Dai
Hubei University of Automotive Technology
Chengzhi Lyu
chengzhi.lyu@huat.edu.cn

Hubei University of Automotive Technology
Hong He
Hubei University of Automotive Technology
Pan Wang
Hubei University of Automotive Technology
Lei Zhang
Hubei University of Automotive Technology

Research Article

Keywords: Monocular depth estimation, self-supervised learning, Diffusion, Discriminative prior

Posted Date: October 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8440557/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

GeoDiffusion:Enhancing Self-Supervised Monocular Depth Estimation in Dynamic Scenes

Fang Dai[1], Chengzhi Lyu[1*], Hong He[1], Pan Wang[1], Lei Zhang[1]

[1]Institute of Automotive Engineers, Hubei University of Automotive Technology, 167 Checheng West Road, Shiyan City, 442000, Hubei, China.

*Corresponding author(s). E-mail(s): chengzhi.lyu@huat.edu.cn;
Contributing authors: 202411434@huat.edu.cn;

Abstract

Contemporary approaches to monocular depth estimation, predominantly relying on self-supervision, typically presuppose static environments; this assumption, however, is often invalid in real-world scenarios such as autonomous driving. In these dynamic contexts, inaccuracies in pose estimation, motion blur, and occlusions arise, disrupting photometric consistency and impairing depth accuracy. To address these challenges, we propose the GeoDiffusion framework, which employs diffusion models for self-supervised depth estimation. Initially, pseudo-ground-truth depth maps are generated using a pretrained Lite-Mono-8M teacher model. Subsequently, a multiview consistency module removes unstable regions and enhances pseudo-label reliability. Following this, a conditional depth diffusion process integrates multiscale image features , progressively refining depth predictions through multi-stage denoising. In addition, a discriminative prior enhancement module ensures geometric consistency and structural fidelity. The proposed method attains accurate monocular depth estimation in dynamic scenes, as shown by results on Cityscapes and KITTI benchmarks.During inference, it uses only the student model, removing the need for ground-truth depth labels or auxiliary networks.

Keywords: Monocular depth estimation, self-supervised learning, Diffusion, Discriminative prior

1 Introduction

Understanding 3D spaces is vital for autonomous driving tasks such as path planning and obstacle detection. Traditional methods like LiDAR and stereo cameras are accurate but costly or calibration-dependent. Monocular depth estimation[25], using a single RGB camera, offers a cheaper and more flexible alternative, divided into supervised and self-supervised approaches.

Supervised methods achieve high accuracy using real depth labels, but such data are costly and limited. Self-supervised methods avoid labels via photometric consistency, assuming static scenes. However, object or camera motion, blur, and occlusion break this assumption, degrading performance in dynamic environments (Fig. 1).

To handle dynamic scenes, previous work combine auxiliary tasks such as semantic[31] or motion segmentation[29], which introduce implicit supervision despite self-supervised training. Semantic segmentation requires costly annotations, while motion segmentation relies on precise trajectory modeling. Diffusion models[8], with strong generative and denoising capacities, better address dynamics, occlusion, and viewpoint changes. However, their use in self-supervised autonomous driving remains underexplored.

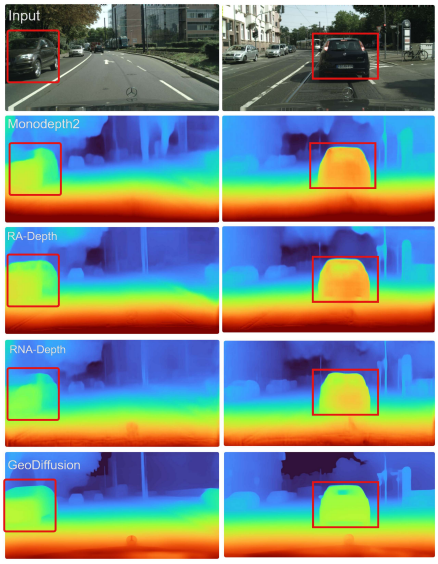


Fig. 1 Depth estimation in dynamic scenes, where moving objects (red boxes) appear, challenges benchmark methods Monodepth2[13], RA-Depth[16], and DNA-Depth[31], unlike our GeoDiffusion approach..

We propose GeoDiffusion, a diffusion-based self-supervised framework for monocular depth estimation in dynamic driving scenes. A pre-trained Lite-Mono-8M teacher generates multi-scale pseudo-depth maps refined by a multiview consistency module to remove unreliable regions from motion or occlusion. The student network performs forward diffusion with noise and learns depth reconstruction via multistage denoising guided by image features . a discriminative prior enhancement module further preserves geometry. Training is fully self-supervised, optimized through knowledge distillation, photometric loss, and diffusion trajectory alignment. After training, the teacher is discarded, allowing a lightweight student to be used for inference.

Our main contributions are summarized as follows.

•We propose GeoDiffusion, a self-supervised monocular depth framework using diffusion to reduce errors in dynamic scenes.

•We design a robust pseudo-depth generation via a pre-trained teacher and multiview consistency verification, enabling label-free supervision.

•We introduce discriminative prior injection to improve geometric and structural fidelity during diffusion.

•Experiments on Cityscapes and KITTI show state-of-the-art accuracy and efficiency, surpassing existing methods.

2 Related Work

2.1 Self-Supervised Monocular Depth Estimation

Self-supervised depth learning avoids ground-truth annotations via photometric reconstruction. Zhou et al.[35] introduced view synthesis with explainability masks; later work added rigid optical flow, automasking and per-pixel minimal loss[13], geometric consistency[3], domain separation[22], wavelet decomposition, edge-aware disparity[18], symmetric encoder-decoder[14], self-distillation[36], disparity offsets[9], hybrid fusion[20], and surface/semantic priors.

In autonomous driving, moving objects violate photometric consistency. Some methods use semantic segmentation (SAFFNet[6], FSRE-Depth[19], IC-MHA[24]), while others explicitly model motion (Struct2Depth[4], DynamicDepth[10], Dynamo-Depth[29]). However, real-world dynamics—rapid motion, occlusion, and lighting changes—remain challenging.

2.2 Diffusion Models for Monocular Depth Estimation

Diffusion models have achieved remarkable results in image synthesis[8], semantic segmentation[30], and depth estimation[28], reconstructing structured information via iterative denoising. In monocular depth estimation, Saxena et al.[32] first applied diffusion in supervised learning to improve structured depth maps. Self-supervised diffusion frameworks include Shao et al.[27] with pseudo-label guidance and SfMDiffusion for endoscopic videos.

However, most methods focus on static or low-dynamic scenes, limiting applicability to autonomous driving with fast motion, occlusion, and lighting changes. To address this, we propose GeoDiffusion, which combines diffusion with self-supervised learning using teacher-generated pseudo-depth maps, multiview consistency, and discriminative priors to guide stepwise denoising,

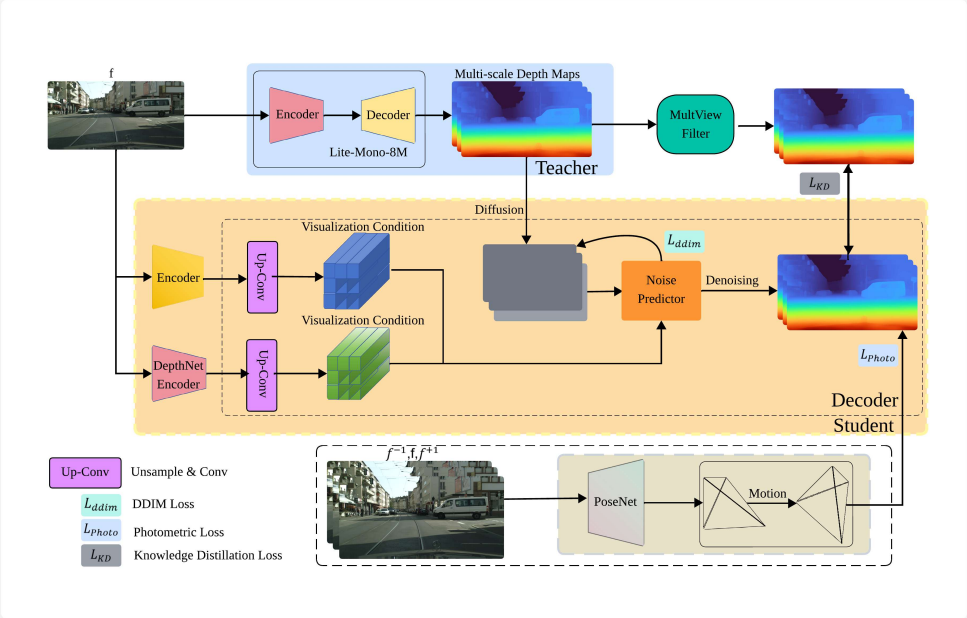


Fig. 2 GeoDiffusion generates cross-view consistent pseudo-depth via a pretrained teacher, while a dual-branch diffusion student uses DDIM denoising, knowledge distillation, and photometric loss to improve depth estimation.

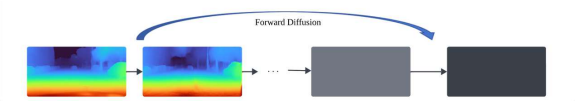


Fig. 3 Forward diffusion adds noise progressively, formineg the basis for GeoDiffusion's reverse denoising.

achieving accurate depth estimation in dynamic driving without ground-truth annotations.

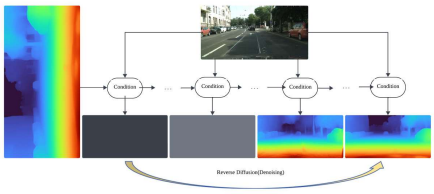


Fig. 4 Denoising leverages visual features and discriminative priors to improve depth accuracy, consistency, and robustness in dynamic scenes.

3 Method

GeoDiffusion is designed to improve robustness in self-supervised monocular depth estimation under dynamic scenes. Its architecture (Fig. 2) includes teacher, student, and pose modules, the student being used inference. Key components: multi-view pseudo-depth, sequential denoising, guiding priors, and structured training losses.

3.1 Network Architecture

The teacher module adopts a lightweight self-supervised monocular depth estimation network, Lite-Mono-8M, as the teacher model. Multi-view consistency filters unreliable pixels (due to motion, occlusion, or reflections) to create robust pseudo-depth maps for diffusion and as discriminative priors.

The student module has two paths: a Lite-Mono DepthNet encoder extracts multiscale visual features guiding denoising, and a pre-trained Lite-Mono-8M encoder provides structural priors. Features are fused through learnable attention weights a_t, modulating contributions per denoising step $\mathcal{T}$. A compact noise predictor (embedding + three conv layers) balances multi-scale features for noise estimation.

Pose estimation uses a pre-trained ResNet-18 encoder with a four-layer conv decoder, taking two consecutive frames and predicting the relative pose of 6-DOF .

3.2 Robust Pseudo-depth Construction

In self-supervised monocular depth estimation, the teacher module provides pseudo-ground-truth depth maps. Using a pre-trained Lite-Mono-8M model, it generates multiscale predictions refined via multiview consistency, which reprojects depth across frames and filters pixels with high variance. The resulting maps supervise the student module. Fig. 3 illustrates the student's forward diffusion process, with its mathematical formulation detailed below.

$$\mathrm{q}(\mathcal{D}_{\mathcal{T}} \mid \mathcal{D}_{\text{pseudo}}) = \mathcal{N}\big(\mathcal{D}_{\mathcal{T}} \mid \sqrt{\alpha_{\mathcal{T}}}\mathcal{D}_{\mathcal{T}},\ (1-\alpha_{\mathcal{T}})\mathcal{I}\big) \quad (1)$$

A consistency filter $\phi(p)$ marks a pixel as consistent ($\phi(p) = 1$) if the standard deviation of its reprojected depth across views is below a threshold δ, and inconsistent ($\phi(p) = 0$) otherwise. This filter guides the loss of knowledge distillation.

$$\mathcal{L}_{KD} = \sum_{p} \phi(p) \odot \big(D_t(p) - D_{\text{pseudo}}(p)\big) \quad (2)$$

Regions affected by specular reflections, fast motion, or occlusions are excluded, enhancing the reliability and robustness of pseudo-labels for student training during diffusion.

3.3 Conditional Diffusion Depth Estimation

In GeoDiffusion, monocular depth estimation is formulated as an iterative denoising task. The student module predicts the depth map of a target frame using multiscale encoder features, guided by teacher-provided depth conditions (Fig. 4). Denoising follows a reverse diffusion process through discrete stages T.

$$p_\theta(\mathcal{D}_{\mathcal{T}-1} \mid \mathcal{D}_{\mathcal{T}}, c) = \mathcal{N}\big(\mathcal{D}_{\mathcal{T}-1} \mid \mu_\theta(\mathcal{D}_{\mathcal{T}}, \mathcal{T}, c), \sigma^2_{\mathcal{T}}\big) \quad (3)$$

Here, c represents conditional information from teacher depth predictions and multiscale features. The noise prediction network outputs μ_θ using $D_{\mathcal{T}}$, $\mathcal{T}$, and c, while $\sigma^2_{\mathcal{T}}$ follows a scheduled decay to progressively reduce noise and refine depth predictions.

3.4 Discriminative Prior Enhancement

To improve student denoising, the depth derived from the teacher D_{teacher}[2] serves as a discriminative prior, providing dynamic guidance during iterative denoising by adjusting the conditional probability distribution.

$$p_\theta(\mathcal{D}_{\mathcal{T}-1} \mid \mathcal{D}_{\mathcal{T}}, D_{\text{teacher}}) = \mathcal{N}\big(\mathcal{D}_{\mathcal{T}-1} \mid \mu_\theta(\mathcal{D}_{\mathcal{T}}, \mathcal{T}, D_{\text{teacher}}), \sigma^2_{\mathcal{T}}\big) \quad (4)$$

The mean function μ_θ controls how the priors provided by the teacher guide the denoising at each step $\mathcal{T}$. The student incrementally refines the predictions by combining $D_{\mathcal{T}}$ with teacher cues D_{teacher}, efficiently transferring knowledge while allowing deviations governed by learnable parameters θ.

3.5 Loss Formulation

GeoDiffusion training combines photometric consistency L_{ph}, teacher-based knowledge distillation, and DDIM losses with discriminative priors.

L_{ph} evaluates the match between the synthesized $f_{s\to t}$ (from f_s and $D_{\mathcal{T}}$) and the target frame using SSIM and $|\mathbf{L}|_1$.

$$L_{ph} = \sum_{p} \Bigg[\kappa \frac{1 - \mathrm{SIM}(f_t(p), f_{s\to t}(p))}{2} + (1-\kappa)\|f_t(p) - f_{s\to t}(p)\|_1\Bigg] \quad (5)$$

SSIM preserves structural integrity and edge alignment, while $|\mathbf{L}|_1$ improves depth accuracy in low-texture regions, improving robustness in various driving scenarios.

The loss of knowledge distillation L_{KD} guides the student to align the predicted depths $\mathcal{T}$ with the pseudo-ground truth of the teacher D_{pseudo}, transferring the knowledge of the teacher to improve the accuracy of the estimation.

$$\mathcal{L}_{KD} = \sum_{p} \phi(p) \odot \big(D_{\mathcal{T}}(p) - D_{\text{pseudo}}(p)\big) \quad (6)$$

DDIM loss L_{DDIM} enforces coherence between student predictions and the expected diffusion progression, ensuring that depth maps follow the

intended denoising trajectory.

$$\mathcal{L}_{\text{DDIM}} = \sum_{\mathcal{T}} \|D_{\mathcal{T}} - \hat{D}_{\mathcal{T}}\|^2 (7)$$

Here, $\hat{D}_{\mathcal{T}}$ denotes the depth predicted in the denoising step $\mathcal{T}$. The quadratic loss penalizes deviations from the intended trajectory, promoting stability and gradual refinement.

The overall loss is a weighted sum of the photometric consistency, the distillation of knowledge, and the DDIM components.

$$\mathcal{L}_{\text{total}} = \lambda_1 L_{ph} + \lambda_2 L_{KD} + \lambda_3 L_{\text{DDIM}} (8)$$

In experiments, weighting coefficients are set as $\lambda_1 = \lambda_2 = \lambda_3 = 1$. Following [12], photometric loss uses a dual-source frame strategy, selecting the frame with lower loss to improve the robustness against occlusions.

4 Experiments

4.1 Datasets and Evaluation Metrics

4.1.1 Datassets

The experiments use Cityscapes[7] and KITTI[11] . Cityscapes includes many moving objects with 69,730 training images and 1,525 evaluation images; depths capped at 100m, resized to 1232×336. KITTI has 39,810 training images, 4,424 validation images, and 697 test images, after Eigen split and Zhou et al. preprocessing; depths also capped at 100 m, resized to 1232×336. Testing uses single-frame depth estimation.

4.1.2 Metrics

Seven standard metrics[13] are used: error-based (AbsRel, SqRel, RMSE, RMSElog; lower better) and accuracy-based ($\delta < 1.25$, 1.25^2, 1.25^3; higher better).

Absolute Relative Error (AbsRel):

$$\text{AbsRel} = \frac{1}{N} \sum_{i=1}^{N} \frac{|D_i - \hat{D}_i|}{D_i} (9)$$

Square Relative Error (Sq Rel):

$$\text{SqRel} = \frac{1}{N} \sum_{i=1}^{N} \frac{(D_i - \hat{D}_i)^2}{D_i} (10)$$

Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(D_i - \hat{D}_i\right)^2} (11)$$

Root Mean Square Logarithmic Error (RMSE log):

$$\text{RMSElog} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\log D_i - \log \hat{D}_i\right)^2} (12)$$

Accuracy with Threshold ($\delta < \text{th}$):

$$\% \text{ of } D_i \text{ such that } \max\left(\frac{\hat{D}_i}{D_i}, \frac{D_i}{\hat{D}_i}\right) = \delta < \text{th},$$
$$\text{th} = 1.25, 1.25^2, 1.25^3 \quad (13)$$

D_i and $\hat{D}_i$ are ground-truth and predicted depths for pixel i; N is the total valid pixels.

4.2 Implementation Details

Experiments use PyTorch 2.4.1 on Ubuntu 20.04 with CUDA 12.4 and an RTX 3090 GPU. AdamW (weight decay $1e^{-2}$), batch size 12, learning rate $1e^{-4}$ with cosine annealing. Diffusion steps tested (500–3000); 2000 chosen for quality–efficiency balance. Fast sampling reduces inference steps to 20 with similar results.

4.3 Comparison to Prior State-of-the-Arts

4.3.1 Analysis of the Cityscapes Dataset

Table 1 compares GeoDiffusion with leading methods on Cityscapes. GeoDiffusion achieves SqRel 0.905 (vs. RA-Depth 0.985), RMSE 5.750 (vs. 5.791), and AbsRel 0.105 (vs. MonoViT 0.115), outperforming prior self-supervised models without semantic or motion labels. Bar charts and Fig.6 visualize these improvements. Qualitative

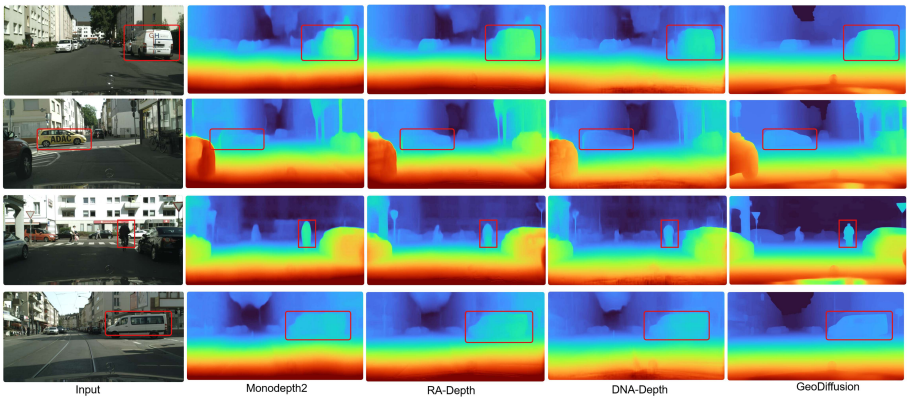


Fig. 5 Qualitative results on Cityscapes show GeoDiffusion accurately predicts depth, particularly for dynamic objects highlighted by red boxes.

Table 1 Comparison with state-of-the-art methods on the Cityscapes dataset.

Methods	Data	AbsRel↓	SqRel↓	RMSE↓	RMSE log↓	$\delta <$ 1.25↑	$\delta <$ 1.25^2↑	$\delta <$ 1.25^3↑
Struct2Depth[4]	M+Sem	0.145	1.737	7.280	0.205	0.813	0.942	0.976
Monodepth2[13]	M	0.129	1.569	6.876	0.187	0.849	0.957	0.983
HR-Depth[23]	M	0.124	1.239	6.726	0.184	0.845	0.958	0.987
Li et al.[21]	M+Sem	0.119	1.290	6.980	0.190	0.846	0.953	0.982
BRNet[15]	M	0.119	1.101	6.723	0.179	0.848	0.962	0.990
MonoVit[34]	M	0.115	1.118	6.339	0.171	0.867	0.967	0.990
RA-Depth[16]	M	0.111	0.985	5.791	0.162	0.882	0.967	0.990
SD-SSMDE[26]	M	0.110	0.988	5.953	0.165	0.876	**0.970**	**0.991**
MonoFormer[1]	M	0.120	1.075	6.405	0.175	0.853	0.966	**0.991**
Freq-aware-depth[5]	M	0.125	1.356	6.618	-	0.856	-	-
Lite-Mono[33]	M	0.123	1.324	6.519	0.181	0.858	0.962	0.987
Dynamo-Depth[29]	M	0.116	1.032	6.089	0.169	0.868	0.964	0.989
DaCCN[21]	M	0.113	1.380	6.305	-	**0.888**	-	-
DNA-Depth[31]	M	0.116	1.209	6.359	0.171	0.870	0.967	0.989
GeoDiffusion(ours)	M	**0.105**	**0.905**	**5.750**	**0.155**	0.882	**0.970**	**0.991**

Note: The best performance is highlighted in bold, and the second-best is underlined. M: trained with monocular video. SEM: trained with semantic segmentation labels.

results (Fig. 5) show GeoDiffusion provides more accurate depth, especially for dynamic objects (red boxes).

4.3.2 Analysis of the KITTI Dataset

Table 2 compares GeoDiffusion with recent methods on KITTI. GeoDiffusion achieves sqRel 0.609 (vs. 0.632) and RMSE 4.180 (vs. 4.416), outperforming RA-Depth and ViT-based models. Fig. 8 shows bar charts, and Fig. 7 presents qualitative depth predictions, highlighting robust performance in dynamic driving scenes.

4.3.3 Complexity and speed evaluation

To assess the practicality of our proposed technique, we compared the sizes of different models. GeoDiffusion includes approximately 7.6 million parameters, notably less than the 9.98 million of RA-Depth, about 8 million of DNA-Depth, and 14.3 million of Monodepth2, highlighting its compactness. Regarding inference speed, GeoDiffusion operates at around 9.2 milliseconds per frame on an RTX 3090, demonstrating its promising capability for real-time applications.

Table 2 Comparison with state-of-the-art methods on the KITTI dataset.

Methods	Data	AbsRel↓	SqRel↓	RMSE↓	RMSE log↓	$\delta <$ 1.25↑	$\delta <$ 1.25^2↑	$\delta <$ 1.25^3↑
Monodepth2[13]	M+Sem	0.115	0.903	4.863	0.193	0.877	0.959	0.981
SAFENet[6]	M	0.112	0.788	4.582	0.187	0.878	0.963	0.983
PackNet-SfM[14]	M	0.111	0.785	4.601	0.189	0.878	0.960	0.982
HR-Depth[23]	M+Sem	0.109	0.792	4.632	0.185	0.884	0.962	0.983
FSRE-Depth[19]	M	0.102	0.675	4.393	0.178	0.893	0.966	0.984
BR-Net[15]	M	0.105	0.698	4.462	0.179	0.890	0.965	0.984
SD-SSMDE[26]	M	0.100	0.661	4.264	0.172	0.896	0.967	**0.985**
MonoVit[34]	M	0.099	0.708	4.372	0.175	0.900	0.967	0.984
RA-Dpeth[16]	M	0.096	0.632	4.216	0.171	0.903	0.968	**0.985**
MonoFormer[1]	M	0.104	0.846	4.580	0.183	0.891	0.962	0.982
GPGS-Depth[17]	M	0.116	0.842	4.708	0.190	0.876	0.961	0.982
Dynamo-Depth[29]	M	0.112	0.758	4.505	0.183	0.873	0.959	0.984
freq-aware-depth[5]	M	0.105	0.711	4.452	0.181	0.886	0.964	0.984
Lite-Mono[33]	M	0.101	0.729	4.454	0.178	0.897	0.965	0.983
DaCCN[21]	M	0.099	0.661	4.316	0.173	0.897	0.967	**0.985**
DNA-Depth[31]	M	0.102	0.757	4.493	0.178	0.896	0.965	0.984
GeoDiffusion(ours)	M	**0.090**	**0.609**	**4.180**	**0.170**	**0.905**	**0.969**	**0.985**

Note: The best performance is highlighted in bold, and the second-best is underlined. M: trained with monocular video. SEM: trained with semantic segmentation labels.

Table 3 Ablation study on the Cityscapes and KITTI datasets.

ID	RPC	DPE	LKD	LPH	Abs Rel↓	RMSE↓	$\delta < 1.25$ ↑
1	√				0.133/0.104	6.592/4.91	0.873/0.889
2	√	√			0.127/0.102	5.981/4.721	0.876/0.891
3	√	√	√		0.107/0.099	5.679/4.611	0.880/0.899
4	√	√		√	0.107/0.094	5.730/4.399	0.873/0.900
5	√	√	√	√	**0.105**/**0.090**	**5.570**/**4.180**	**0.882**/**0.905**

Note: RPC: Robust Pseudo-depth Construction; DPE: Discriminative Prior Enhancement; LKD: Knowledge Distillation Loss; LPH: Photometric Loss. Bold values indicate the best performance.

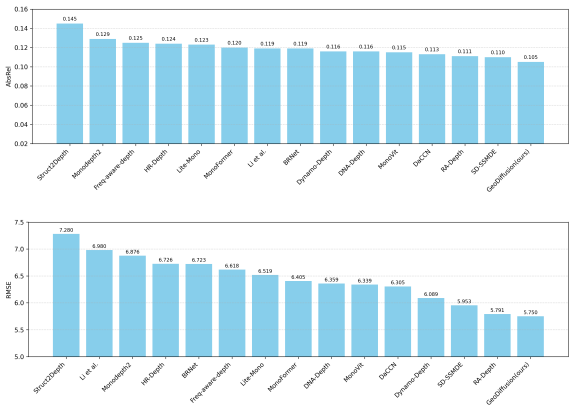


Fig. 6 Bar charts of AbsRel and RMSE on Cityscapes, with methods ordered by decreasing error.

4.4 Ablation Studies

4.4.1 Functional Analysis of Key Modules

In order to thoroughly explain the role of each component within the GeoDiffusion framework, a series of detailed ablation studies were conducted using the Cityscapes and KITTI datasets, as specified in Table 3.

As shown in Table 3, the integration of key components leads to a substantial enhancement in model performance. In the Cityscapes dataset, implementing the DPE, LKD, and LPH modules systematically lowers Abs Rel from 0.133 to 0.105

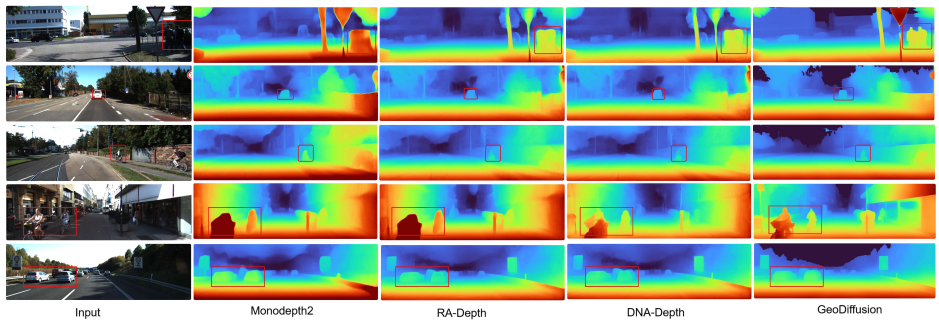


Fig. 7 KITTI qualitative results: GeoDiffusion predicts depth accurately for moving vehicles and pedestrians (red boxes).

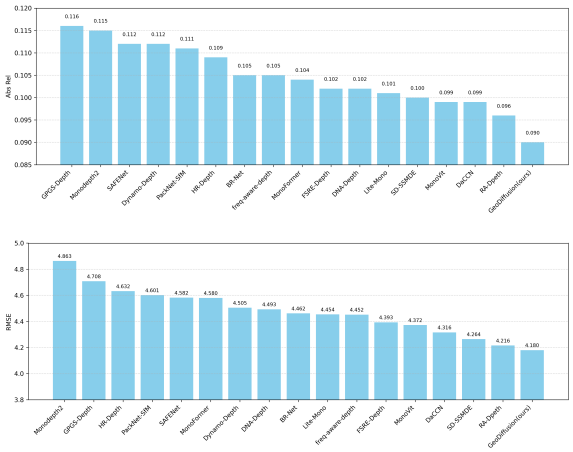


Fig. 8 Bar chart visualizations of Abs Rel and RMSE on the KITTI dataset. The methods are systematically organized in descending order according to their corresponding error values.

and reduces RMSE from 6.592 to 5.570. Similarly, in the KITTI dataset, Abs Rel decreases from 0.104 to 0.090, and RMSE from 4.910 to 4.180. The color gradients clearly show this trend, highlighting the positive impact of each module on overall performance. This pattern confirms the efficacy and resilience of the multicomponent integration strategy across different evaluation metrics.

5 Conclusions

This study presents GeoDiffusion, a self-supervised framework for estimating monocular depth in dynamic scenes. Leveraging pseudo-depth maps from the pre-trained Lite-Mono-8M model, it guides the diffusion-based denoising process to enhance depth modeling in dynamic environments. Through knowledge distillation and photometric consistency loss, GeoDiffusion achieves superior performance on Cityscapes and KITTI, demonstrating strong potential to advance 3D perception. However, its dependence on pseudo-ground truths may introduce teacher-model biases under challenging lighting or weather, affecting generalization. Future work will focus on improving robustness to dynamic objects and developing stronger self-supervised strategies to reduce pseudo-label reliance, thereby enhancing adaptability in real-world autonomous driving.

Acknowledgements. This research was financially supported by the Hubei Province Natural Science Foundation (Grant No. 2024AFD116), the Key Project under the Leadership initiative at Hubei University of Automotive Technology (Grant No. 2024JBB08), the Hubei University of Automotive Technology Doctoral Scientific Research Foundation (Grant No. BK202102), and the Hubei Provincial International Science and Technology Cooperation Program (Grant No. 2025EHA024).

Author Contribution. Fang Dai: Writing an original draft, validation, project administration, methodology, investigation, formal analysis, conceptualization.Hong He: Software, Writing – review & editing.Zhiwei Liu: Resources,Data curation.Lei Zhang: Visualization and Validation.Pan Wang: Formal analysis, Data curation.

Data Availability. The data presented in this study are available on request from the corresponding author. The data is not publicly available due to privacy.

Funding. This work was partially supported by the Hubei Provincial International Science and Technology Cooperation Program (Grant No. 2025EHA024), the Hubei Province Natural Science Foundation, China (Grant No. 2024AFD116), the Key Project led by the Project Leader under the Open Competition for Leadership initiative at the Hubei University of Automotive Technology (Grant No. 2024JBB08), and the Hubei University of Automotive Technology Doctoral Scientific Research Foundation (Grant No. BK202102).

Declarations

Conflict of interests. The authors declare no conflict of interest.

References

[1] Bae J, Moon S, Im S (2023) Deep digging into the generalization of self-supervised monocular depth estimation. In: Proceedings of the AAAI conference on artificial intelligence, pp 187–196

[2] Bergmann P, Fauser M, Sattlegger D, et al (2020) Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4183–4192

[3] Bian J, Li Z, Wang N, et al (2019) Unsupervised scale-consistent depth and ego-motion learning from monocular video. Advances in neural information processing systems 32

[4] Casser V, Pirk S, Mahjourian R, et al (2019) Unsupervised monocular depth and ego-motion learning with structure and semantics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 0–0

[5] Chen X, Li TH, Zhang R, et al (2023) Frequency-aware self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 5808–5817

[6] Choi J, Jung D, Lee D, et al (2020) Safenet: Self-supervised monocular depth estimation with semantic-aware feature extraction. arXiv preprint arXiv:201002893

[7] Cordts M, Omran M, Ramos S, et al (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3213–3223

[8] Croitoru FA, Hondru V, Ionescu RT, et al (2023) Diffusion models in vision: A survey. IEEE transactions on pattern analysis and machine intelligence 45(9):10850–10869

[9] Dai H, Zhang X, Zhao Y, et al (2021) Adaptive disparity candidates prediction network for efficient real-time stereo matching. IEEE Transactions on Circuits and Systems for Video Technology 32(5):3099–3110

[10] Feng Z, Yang L, Jing L, et al (2022) Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In: European Conference on Computer Vision, Springer, pp 228–244

[11] Geiger A, Lenz P, Stiller C, et al (2013) Vision meets robotics: The kitti dataset. The international journal of robotics research 32(11):1231–1237

[12] Godard C, Mac Aodha O, Brostow GJ (2017) Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 270–279

[13] Godard C, Mac Aodha O, Firman M, et al (2019) Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3828–3838

[14] Guizilini V, Ambrus R, Pillai S, et al (2020) 3d packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2485–2494

[15] Han W, Yin J, Jin X, et al (2022) Brnet: Exploring comprehensive features for monocular depth estimation. In: European Conference on Computer Vision, Springer, pp 586–602

[16] He M, Hui L, Bian Y, et al (2022) Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In: European Conference on Computer Vision, Springer, pp 565–581

[17] Jia S, Pei X, Yao W, et al (2023) Self-supervised depth estimation leveraging global perception and geometric smoothness. IEEE

Transactions on Intelligent Transportation Systems 24(2):1502–1517

[18] Johnston A, Carneiro G (2020) Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition, pp 4756–4765

[19] Jung H, Park E, Yoo S (2021) Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 12642–12652

[20] Lee S, Park J, Sohn KM (2022) Hybrid feature fusion networks for monocular depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(5):2307–2319

[21] Li H, Gordon A, Zhao H, et al (2021) Unsupervised monocular depth learning in dynamic scenes. In: Conference on Robot Learning, PMLR, pp 1908–1917

[22] Liu L, Song X, Wang M, et al (2021) Self-supervised monocular depth estimation for all day images using domain separation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 12737–12746

[23] Lyu X, Liu L, Wang M, et al (2021) Hr-depth: High resolution self-supervised monocular depth estimation. In: Proceedings of the AAAI conference on artificial intelligence, pp 2294–2301

[24] Ma J, Lei X, Liu N, et al (2022) Towards comprehensive representation enhancement in semantics-guided self-supervised monocular depth estimation. In: European Conference on Computer Vision, Springer, pp 304–321

[25] Natan O, Miura J (2022) End-to-end autonomous driving with semantic depth cloud mapping and multi-agent. IEEE Transactions on Intelligent Vehicles 8(1):557–571

[26] Petrovai A, Nedevschi S (2022) Exploiting pseudo labels in a self-supervised learning framework for improved monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1578–1588

[27] Shao J, Liu SJ, Ma Z, et al (2024) Self-supervised monocular depth estimation via pseudo-label guided diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 1384–1394

[28] Song Z, Wang Z, Li B, et al (2025) Depthmaster: Taming diffusion models for monocular depth estimation. arXiv preprint arXiv:250102576

[29] Sun Y, Hariharan B (2023) Dynamo-depth: Fixing unsupervised depth estimation for dynamical scenes. Advances in Neural Information Processing Systems 36:54987–55005

[30] Tan W, Chen S, Yan B (2023) Diffss: Diffusion model for few-shot semantic segmentation. arXiv preprint arXiv:230700773

[31] Wang B, Wang S, Ye D, et al (2024) Deep neighbor layer aggregation for lightweight self-supervised monocular depth estimation. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 4405–4409

[32] Zhang F, Yu Z, Li C, et al (2025) Structure-aware radar-camera depth estimation. arXiv preprint arXiv:250605008

[33] Zhang N, Nex F, Vosselman G, et al (2023) 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, pp 18537–18546

[34] Zhao C, Zhang Y, Poggi M, et al (2022) Monovit: Self-supervised monocular depth estimation with a vision transformer. In: 2022 international conference on 3D vision (3DV), IEEE, pp 668–678

[35] Zhou T, Brown M, Snavely N, et al (2017) Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1851–1858

[36] Zhou Z, Dong Q (2022) Self-distilled feature aggregation for self-supervised monocular depth estimation. In: European Conference on Computer Vision, Springer, pp 709–726