DiffSQL: Leveraging Diffusion Model for Zero-Shot Self-Supervised Monocular Depth Estimation

Heyuan Zheng, **Yunji Liang** *, **Lei Liu** and **Zhiwen Yu**
School of Computer Science, Northwestern Polytechnical University
2019302741@mail.nwpu.edu.cn, liangyunji@nwpu.edu.cn, liu.lei@mail.nwpu.edu.cn, zhiwenyu@nwpu.edu.cn

Abstract

Self-supervised monocular depth estimation has attracted significant attention due to its broad applications in autonomous driving and robotics. Although significant performance improvement has been achieved by learning the relative distance of objects with the introduction of Self Query Layer (SQL), it struggles with zero-shot generalization due to the lack of geometric features and the fixed number of query size. To address these problems, we propose a diffusion-augmented self-supervised depth estimation framework, dubbed *DiffSQL*, to learn the geometric priors for feature augmentation. We also introduce a dynamic self-query layer that implicitly computes the relative distances between objects by adjusting the query size according to the feature distribution. Experimental results on the KITTI dataset show that *DiffSQL* outperforms SQLdepth by **1.03%** in terms of **AbsRel** and **2.79%** in terms of **SqRel**. Furthermore, our experiments demonstrate that *DiffSQL* is superior in zero-shot generalization.

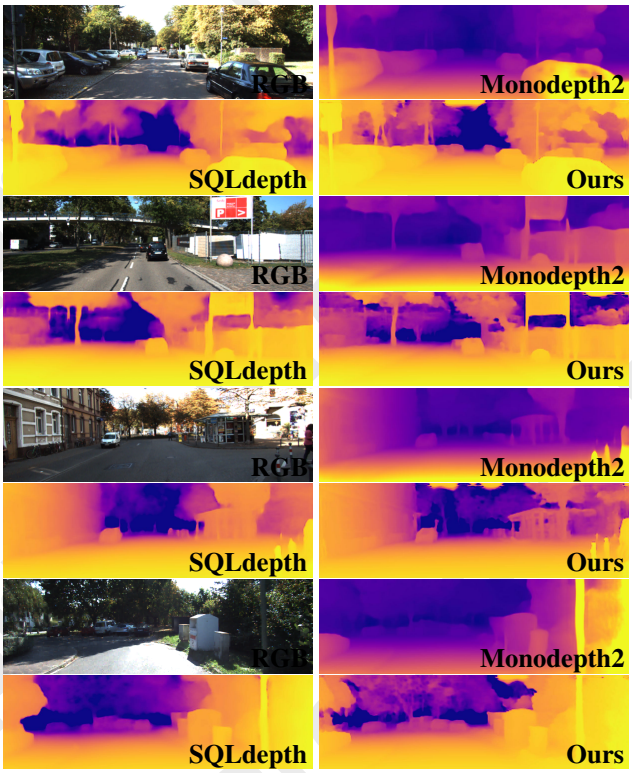


Figure 1: Typical examples of our predictions on images from the KITTI [Geiger *et al.*, 2013] dataset. Compared with Monodepth2 [Godard *et al.*, 2019b] and SQLdepth [Wang *et al.*, 2024], DiffSQL is able to predict depth with more fine-grained details, particularly for thin, small, and distant objects.

1 Introduction

Monocular depth estimation is a fundamental challenge in computer vision, with broad applications in autonomous driving [Geiger *et al.*, 2013], augmented reality [Newcombe *et al.*, 2011], and robotics [Achtelik *et al.*, 2009]. The main objective of this task is to predict the depth of each pixel from a single RGB image. Traditional supervised learning methods rely on sparse ground truth depth data, typically obtained from sensors like LiDAR. However, collecting large amounts of depth data is both costly and time-consuming. Additionally, due to sparse supervision, these methods face difficulties during optimization and struggle with new, unseen scenes due to the lack of zero-shot generalization.

In recent years, self-supervised methods have gained considerable attention due to their ability to eliminate reliance on costly ground truth depth data. SQLdepth [Wang *et al.*, 2024] leverages motion cues and the Self Query Layer (SQL) to infer depth information. However, like many existing methods [Godard *et al.*, 2019b], it uses convolutional neural networks (CNNs) as the backbone for feature extraction, which lack essential spatial geometric features during the construction of the self-query layer. This limitation leads to the neglect of important geometric details in distant and small objects, hindering the model's ability to understand scene struc-

*Corresponding author

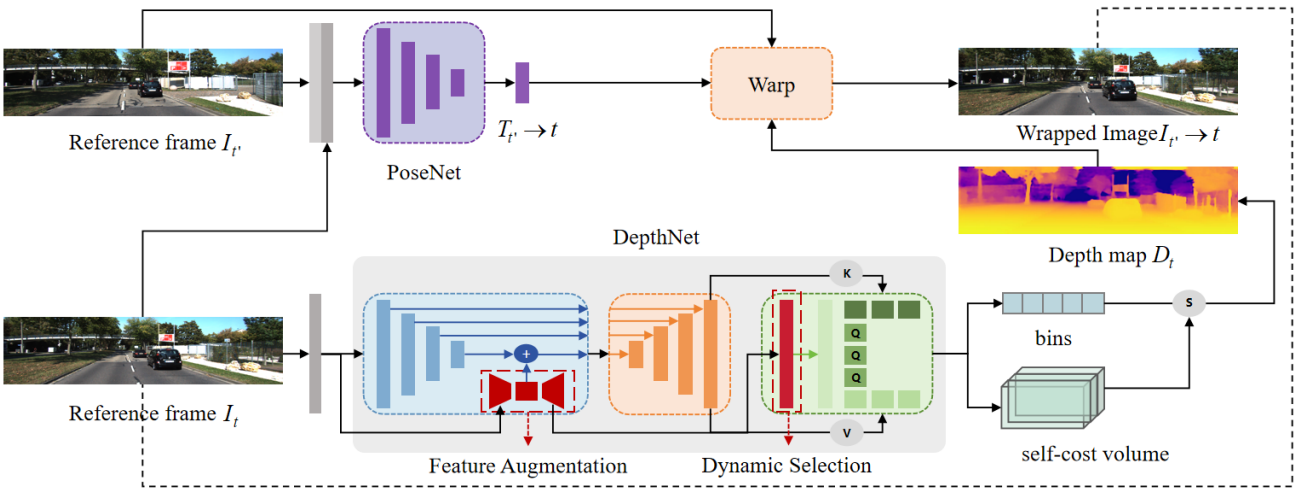


Figure 2: Framework Overview: (1) DepthNet: The system initiates with a hybrid convolution-diffusion feature encoder that processes the input frame I_t to produce multi-scale visual embeddings. These hierarchical features are enhanced by integrating coarse-level diffusion outputs through a Self- Attention mechanism (visualized in Fig. 3) to output the final depth prediction D_t. (2) PoseNet: A conventional pose estimation subnetwork computes the inter-frame transformation matrix $T_{t'\rightarrow t}$ between the target frame I_t and adjacent reference frame $I_{t'}$. This geometric relationship is exclusively utilized during training for view synthesis through differentiable warping operations. (3) Differentiable Image Warping: In this step, pixels from the reference frame I'_t are used to reconstruct the current frame I_t by leveraging the depth map D_t and the relative pose $T_{t'\rightarrow t}$ through a differentiable image warping process [Jaderberg *et al.*, 2015]. The loss function is constructed based on the difference between the warped image $I_{t'\rightarrow t}$ and the source image I_t.

ture. The coarse-grained query objects generated in this manner fail to adequately represent the scene structure, resulting in suboptimal depth map accuracy. Recent studies have demonstrated that diffusion models [Wolleb *et al.*, 2022], such as Stable Diffusion (SD)[Rombach *et al.*, 2021], excel at feature extraction, effectively capturing high-level semantic information and learning geometric features. Motivated by these advancements, we propose DiffSQL, a novel self-supervised monocular depth estimation framework that leverages the powerful semantic extraction capabilities of diffusion models to improve the model's ability to learn and augment features with prior knowledge, thereby enhancing scene understanding.

As shown in Fig. 2, we designed a plug-and-play feature fusion module that integrates texture and semantic features at different scales to improve the model's ability to capture spatial features of distant and small objects. Furthermore, we propose an adaptive self-querying layer that dynamically selects features extracted by the diffusion model using a self-attention mechanism to construct coarse-grained scene object representations. By calculating the correlation between the fused feature map and the coarse-grained object representations, we obtain implicit relative distance information, ultimately resulting in high-precision depth maps.

Our key contributions are as follows:

- We design a plug-and-play diffusion-augmented module that uses the powerful semantic extraction capabilities of diffusion models to complement spatial structure features, enhancing the model's ability to capture distant and small objects.
- We introduce an adaptive self-querying layer that employs attention features extracted by the diffusion model as coarse-grained object representations. By dynamically selecting query objects using a self-attention mechanism, this approach significantly improves depth estimation accuracy, particularly for distant and small objects.
- Through experiments on the KITTI dataset, we show that DiffSQL outperforms existing self-supervised methods in accuracy and efficiency. Zero-shot evaluation on the Make3D dataset further demonstrates the excellent generalization, especially for thin, small and distant objects.

2 Related Works

2.1 Monocular Depth Estimation

Monocular depth estimation (MDE) deduces 3D depth from a single 2D image. The task's ill-posed nature stems from infinite plausible 3D scene representations per image. Current approaches bifurcate into supervised and self-supervised frameworks.

Supervised Depth Estimation. Supervised learning employs ground-truth depth maps as supervisory signals for precise prediction. Eigen [Eigen *et al.*, 2014] pioneered multi-scale CNN architectures for depth estimation. Regression-based approaches [Huynh *et al.*, 2020] predict continuous depth yet suffer convergence instability, whereas classification-based frameworks [Diaz and Marathe, 2019; Fu *et al.*, 2018] enhance optimization stability via discretized depth representation. AdaBins [Bhat *et al.*, 2021] innovatively unifies classification-regression dual pathways. Re-

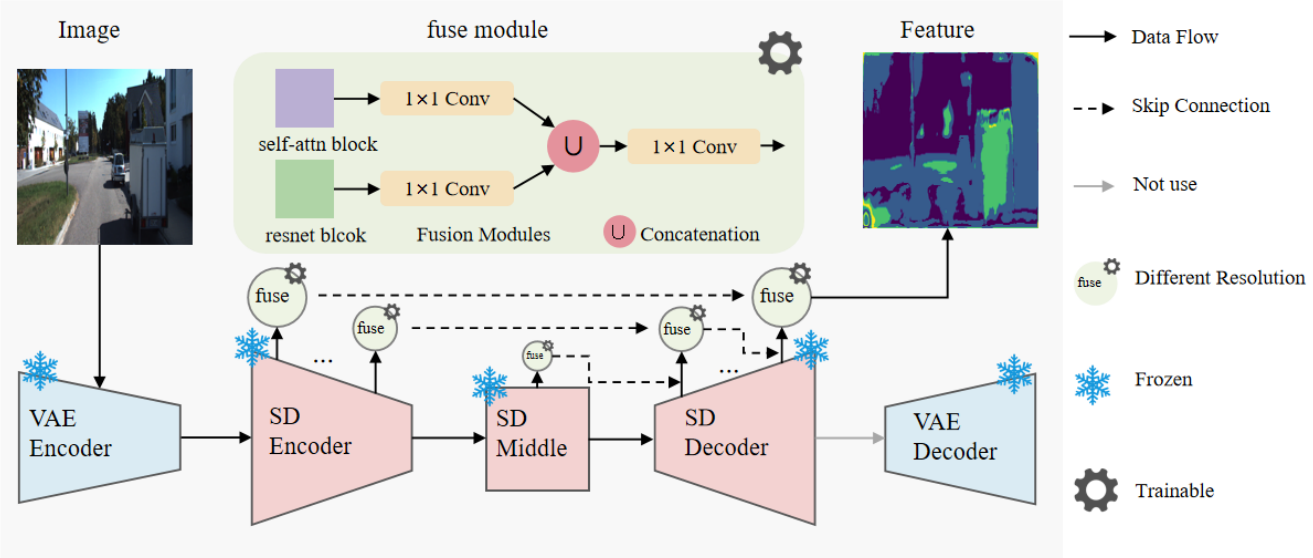


Figure 3: Framework Details of DiffSQL with Diffusion-Augmented Feature Fusion.

cent breakthroughs involve: neural window CRFs optimizing computational efficiency [Yuan *et al.*, 2022], variational constraints strengthening spatial coherence [Liu *et al.*, 2023], and decoupled surface normal/distance estimation [Shao *et al.*, 2023], collectively advancing accuracy and robustness.

Self-Supervised Depth Estimation. Self-supervised approaches circumvent ground-truth dependency by generating supervision from stereo/monocular sequences. Zhou et al. [Zhou *et al.*, 2017] pioneered co-training depth-pose networks with view synthesis losses, incorporating explainability masks for dynamic objects. Subsequent works [Godard *et al.*, 2019a] enhanced robustness via auto-masking and minimum reprojection losses. Garg et al. [Garg *et al.*, 2016] established photometric consistency for stereo pairs, later refined with left-right consistency [Godard *et al.*, 2017]. Cutting-edge developments include continuous disparity prediction [Garg *et al.*, 2020] and hybrid CNN-Transformer architectures [Zhang *et al.*, 2023]. Current methodologies diverge in feature extraction: pure CNN implementations coexist with multi-network integrated architectures, synergistically boosting performance and robustness.

2.2 Diffusion Model

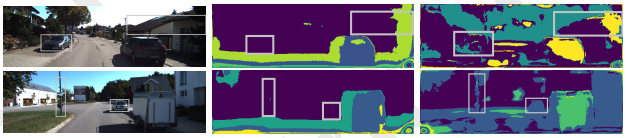
Diffusion models have demonstrated substantial breakthroughs in conditional/unconditional image generation [Ho *et al.*, 2020; Dhariwal and Nichol, 2021]. By modeling complex data distributions through progressive denoising, these models excel in text-to-image synthesis [Koh *et al.*, 2024] and cross-domain image translation [Saharia *et al.*, 2022], achieving superior detail reconstruction. Their hierarchical feature extraction capabilities extend to discriminative tasks including image segmentation [Wolleb *et al.*, 2022] and object detection [Chen *et al.*, 2023], effectively capturing multi-scale semantic patterns. Chen et al. [Chen *et al.*, 2024] innovatively redesigned the tokenizer with linear decay mechanisms, validating diffusion models' potential as universal feature extractors. Leveraging pre-trained Stable Diffusion's cross-modal representations, our work achieves efficient zero-shot monocular depth estimation via single-pass inference, balancing precision and computational efficiency.

3 Methodology

To improve the zero-shot generalization, we propose a diffusion-augmented self-supervised depth estimation framework, dubbed DiffSQL, to learn the geometric priors for feature augmentation and capture the relative distances with dynamic self-cost volume. As shown in Fig. 2, DiffSQL consists of two key components: (1) a diffusion-augmented encoder-decoder to take advantages of pre-trained diffusion models for feature representation; and (2) a dynamic self-query layer that uses a self-attention mechanism to dynamically select coarse-grained query objects to quantify the relative distances between feature distributions.

3.1 Diffusion Encoder for Feature Augmentation

For the feature extraction of images, prior studies mainly rely on convolutional operations and downsampling to learn the latent representation. The typical feature encoders include ResNet. However, existing feature encoders are not informative.During monocular self-supervised training, ResNet demonstrates limitations in capturing geometric features effectively, often neglecting information about distant and small objects, as shown in Fig. 4. Recently, several studies have demonstrated that diffusion-based encoder shows impressive performance for feature extraction [Namekata *et al.*, 2023; Baranchuk *et al.*, 2021]. Accordingly, we conduct comparative studies to investigate the differences of ResNet50 and diffusion models for feature extraction. Specifically, used the k-means algorithm on feature maps extracted by ResNet50 (used in SQLDepth) and our diffusion-



(a) RGB (b) Resnet50 (c) DiffSQL(Ours)

Figure 4: A visual comparison of features extracted from **ResNet50** and our proposed **(DiffSQL)**, using k-means clustering examples (k=8), demonstrates that our approach is more effective to learn the geometric structure of distant and small objects.

augmented feature extractor. The number of cluster centers was set to 8.

Inspired by these observations, we propose a diffusion-augmented encoder-decoder framework that leverages pre-trained diffusion models to enhance feature representation. As shown in Fig. 3, the method extracts multi-scale features from the diffusion model, encompassing its ResNet and self-attention modules. The diffusion model captures global semantic information through its encoder and decoder while preserving local patterns and fine details, achieving an effective balance between global semantics and local features.

The approach obtains multi-scale features from the diffusion model, including its ResNet and self-attention modules. Through its encoder and decoder, the diffusion model captures global semantic information while maintaining local patterns and details, balancing effectively between global semantics and local features.

The input image, with an initial size of (h, w), is processed by the VAE encoder to downsample the feature map to $(h/8, w/8)$. The ResNet module further processes this feature map, maintaining the resolution $(h/8, w/8)$. Concurrently, the self-attention module processes the latent representation at an initial resolution of $(h \times w/64)$, which is subsequently resized to $(h/8, w/8)$. This process generates multi-scale local and global features within the diffusion model, forming the basis for feature fusion.

To effectively integrate the extracted multi-scale features, we designed a feature fusion module. First, a 1×1 convolution is applied to the local features from the ResNet module and the global features from the self-attention module to unify their channel dimensions. These features are then concatenated along the channel dimension, followed by a second 1×1 convolution to reduce channel dimensionality, optimizing the feature representation and minimizing computational overhead. This fusion process is repeated across the encoder, latent, and decoder stages of the diffusion model, with skip connections incorporated to produce the final diffusion-augmented features. This design maximizes the potential of the diffusion model by seamlessly integrating its local and global features, resulting in a more expressive feature representation.

Furthermore, as shown in the red box in Fig. 2, we concatenate features extracted from the SD model with those from the U-Net-CNN backbone at the fourth down-sampling stage along the channel dimension to further enhance feature representation. The fused features are processed through an

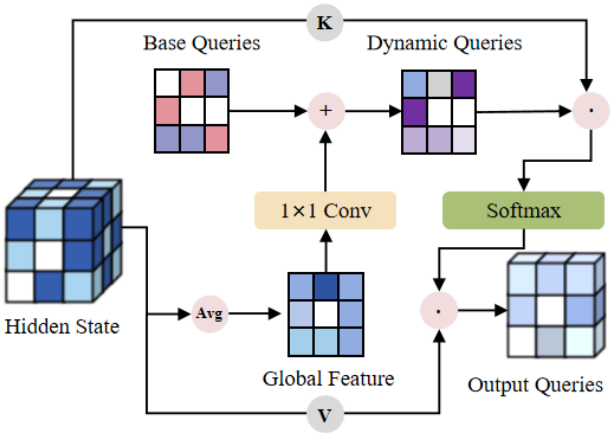


Figure 5: Dynamic query selection process for generating adaptive query vectors from coarse-grained query vectors extracted by the SD model.

up-sampling strategy to generate the network output, resulting in a feature map enriched with spatial geometric details. Figure 4 provides a visual analysis of the feature fusion results.

3.2 Dynamic Self Query Layer

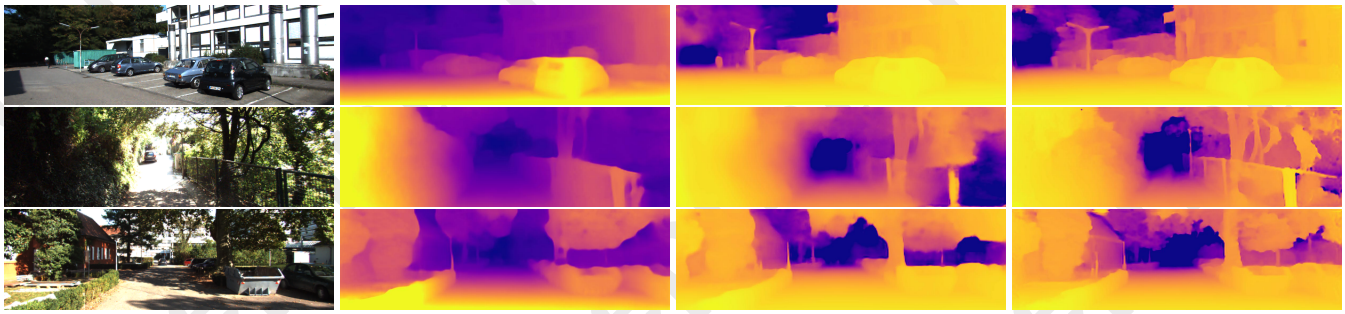
Monocular depth estimation benefits from geometric cues, such as relative distance, to enhance accuracy. A self-cost volume [Wang *et al.*, 2024] to capture relative distances within the same image is an effective approach but incurs a high computational complexity of $\mathcal{O}(h^2 \times w^2)$ when directly applied to high-resolution feature maps. Previous methods address this by using coarse-grained queries to represent objects, reducing complexity. However, when the feature map S fails to capture sufficient semantic or structural details, the effectiveness of coarse-grained queries diminishes, leading to poor model performance.

To overcome this, we replace the self-attention vectors from the Vision Transformer (ViT) with those extracted from the pre-trained Stable Diffusion (SD) model, which excels in capturing global semantic information and contextual relationships. The SD model, trained on diverse datasets, offers rich feature representations that are more suitable for constructing coarse-grained queries. In addition, we introduce a dynamic query selection mechanism, as shown in Fig. 5, to further enhance query flexibility. Given the input feature map $\mathbf{H}_i \in R^{B \times L \times C}$, where B is the batch size, L the sequence length, and C the feature dimension, we compute a global context vector $\mathbf{g}$ through global average pooling, as formulated in Eq. (1):

$$\mathbf{g} = \frac{1}{L} \sum_{j=1}^{L} \mathbf{H}_{i,j}, \quad \mathbf{g} \in R^{B \times C}. \tag{1}$$

This vector is adjusted by a lightweight network (e.g., 1×1 convolution) to produce $\mathbf{g}_{\text{adjusted}}$, as shown in Eq. 2:

$$\mathbf{g}_{\text{adjusted}} = f_{\text{adjust}}(\mathbf{g}), \quad \mathbf{g}_{\text{adjusted}} \in R^{B \times C}. \tag{2}$$



(a) RGB (b) Monodepth2 (c) SQLdepth (d) DiffSQL(Ours)

Figure 6: Additional qualitative results on the KITTI eigen benchmark.

The adjusted vector is added to the base query vectors $\mathbf{Q}_{\text{base}} \in R^{K\times C}$ to generate dynamic queries $\mathbf{Q}_{\text{dynamic}}$ defined in Eq. (3):

$$\mathbf{Q}_{\text{dynamic}} = \mathbf{Q}_{\text{base}} + \mathbf{g}_{\text{adjusted}}, \quad \mathbf{Q}_{\text{dynamic}} \in R^{B\times K\times C}. \tag{3}$$

These dynamic queries are used to compute attention scores with the input hidden state $\mathbf{H}_i$ via dot product according to Eq. (4):

$$\mathbf{A}_{i,j} = \mathbf{Q}^{\top}_{\text{dynamic},i} \cdot \mathbf{H}_{i,j}, \quad \mathbf{A} \in R^{B\times K\times L}. \tag{4}$$

The attention scores are normalized using Softmax, as shown in Eq. (5):

$$\mathbf{A}_{\text{softmax},i,j} = \frac{\exp(\mathbf{A}_{i,j})}{\sum_{j=1}^{L}\exp(\mathbf{A}_{i,j})}, \quad \mathbf{A}_{\text{softmax}} \in R^{B\times K\times L}. \tag{5}$$

The normalized attention scores are then used to compute the output query vectors by performing a weighted sum over the input hidden state, as described in Eq. (6):

$$\mathbf{Q}_{\text{output},i} = \sum_{j=1}^{L} \mathbf{A}_{\text{softmax},i,j} \cdot \mathbf{H}_{i,j}, \quad \mathbf{Q}_{\text{output}} \in R^{B\times K\times C}. \tag{6}$$

Finally, the dynamic query vectors are used to construct the self-cost volume V according to Eq. (7):

$$V_{i,j,k} = \mathbf{Q}^{\top}_{\text{dynamic},i} \cdot \mathbf{S}_{j,k}, \quad \forall\, i \in [1, Q],\ j \in [1, h],\ k \in [1, w]. \tag{7}$$

In our previous work, we proposed a self-cost volume-based method for continuous depth estimation, where depth distributions (depth bins) are redefined as statistical distributions of depth values. This approach leverages softmax and weighted sum operations to aggregate latent depth information and compute the statistical distribution according to Eq. (8), where b represents the statistical depth distribution vector.

$$b = \text{MLP}\left(\sum_{i=1}^{Q}\sum_{(j,k)=(1,1)}^{(h,w)} \text{softmax}(V_i)_{j,k} \cdot S_{j,k}\right), \tag{8}$$

To generate the final depth map, we combine these depth distributions using a probabilistic method. First, a 1×1 convolution maps the self-cost volume V to a D-plane volume, where D matches the depth bin dimension. A plane-wise softmax operation is then applied to obtain the probabilistic map $p_{i,j,k}$ for each plane, as shown in Eq. (9):

$$p_{i,j,k} = \text{softmax}(V)_{i,j,k}, \quad 1 \le i \le Q. \tag{9}$$

The final depth value for each pixel is computed as a probabilistic linear combination of the bin centers in Eq. (10), where $c(b_i)$ represents the center depth of the i-th depth bin in Eq. (11):

$$\tilde{d} = \sum_{i=1}^{N} c(b_i) p_{i,j,k}, \quad 1 \le j \le h,\ 1 \le k \le w, \tag{10}$$

$$c(b_i) = d_{\text{min}} + (d_{\text{max}} - d_{\text{min}})\left(\frac{b_i}{2} + \sum_{j=1}^{i-1} b_j\right). \tag{11}$$

This approach effectively aggregates latent depth information from the self-cost volume to enable precise depth estimation, particularly for continuous depth prediction tasks.

3.3 Loss Functions

Objective Functions. Following the methodologies of [Godard *et al.*, 2019a]. and their extensions, we utilize the standard photometric error p_e as the primary objective function. This combines L1 loss with SSIM to evaluate photometric consistency, as shown in Eq. (12), where α is a balancing parameter.

$$p_e(I_a, I_b) = \alpha \cdot \frac{1}{2}(1 - \text{SSIM}(I_a, I_b)) + (1-\alpha)\|I_a - I_b\|_1, \tag{12}$$

To regularize depth in textureless regions, we include an edge-aware smoothness loss in Eq. (13):

$$L_s = |\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^*| e^{-|\partial_y I_t|} \tag{13}$$

Method	Train	Test	H×W	AbsRel↓	SqRel↓	RMSE↓	RMSElog↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
PackNet-SfM[Guizilini *et al.*, 2020]	M	1	640×192	0.111	0.787	4.601	0.189	0.878	0.960	0.982
HR-Depth[Lyu *et al.*, 2021]	MS	1	640×192	0.107	0.785	4.612	0.185	0.887	0.962	0.982
MonoDepth2[Godard *et al.*, 2019b]	MS	1	640×192	0.106	0.861	4.699	0.185	0.889	0.962	0.982
ManyDepth[Watson *et al.*, 2021]	M	1	640×192	0.099	0.773	4.434	0.178	0.895	0.965	0.983
MonoDiffusion[Shao *et al.*, 2024]	M	1	640×192	0.099	0.702	4.385	0.176	0.899	0.965	0.983
SQLdepth[Wang *et al.*, 2024]	M	1	640×192	0.097	0.718	4.376	0.172	0.900	**0.966**	0.983
Ours(DiffSQL)	M	1	640×192	**0.096**	**0.698**	**4.338**	**0.171**	**0.902**	**0.966**	**0.984**
Ours(DiffSQL)	MS	1	640×192	**0.094**	**0.684**	**4.306**	**0.169**	**0.902**	**0.967**	**0.984**

Table 1: Performance comparison on the KITTI [Geiger *et al.*, 2013] Eigen benchmark. M denotes monocular training, MS combines monocular videos with stereo pairs. Testing uses single-frame input (marked 1). Best results in **bold**, with self-supervised methods applying [Eigen and Fergus, 2015]'s median scaling for depth scale recovery.

Method	Train	Test	H×W	AbsRel↓	SqRel↓	RMSE↓	RMSElog↓	$\delta < 1.25$↑	$\delta < 1.25^2$↑	$\delta < 1.25^3$↑
PackNet-SfM[Guizilini *et al.*, 2020]	M	1	640×192	0.078	0.420	3.485	0.121	0.931	0.986	0.996
MonoDepth2[Godard *et al.*, 2019b]	MS	1	640×192	0.080	0.466	3.681	0.127	0.926	0.985	0.995
CADepthNet[Yan *et al.*, 2021]	M	1	640×192	0.080	0.442	3.639	0.124	0.927	0.986	0.996
AQUANet[Bello *et al.*, 2024]	M	1	640×192	0.073	0.374	3.572	0.115	0.935	0.985	0.996
Dynamic Depth[Feng *et al.*, 2022]	M	2(-1,0)	640×192	0.068	0.362	3.454	0.111	0.943	0.988	0.996
SQLdepth[Wang *et al.*, 2024]	M	1	640×192	0.068	0.359	3.347	**0.105**	0.944	**0.989**	**0.997**
Ours (DiffSQL)	M	1	640×192	**0.067**	**0.343**	**3.312**	0.107	**0.946**	**0.989**	**0.997**
Ours (DiffSQL)	MS	1	640×192	**0.065**	**0.338**	**3.259**	**0.105**	**0.947**	**0.990**	**0.998**

Table 2: Performance comparison using KITTI improved ground truth from [Uhrig *et al.*, 2017].

Masking Strategy. To address challenges in self-supervised depth estimation, we propose an auto-masking strategy that filters out stationary pixels and low-texture regions by leveraging temporal photometric differences. The mask μ is defined as Eq. 14.

$$\mu = \left[\min_{t'} p_e(I_t, I_{t'\to t}) < \min_{t'} p_e(I_t, I_{t'}) \right] \tag{14}$$

Final Training Loss. The total loss function defined in Eq. (15) combines the photometric loss, the smoothness loss, and the auto-masking strategy. Here λ balances the two loss terms.

$$L = \mu \cdot L_{\text{photo}} + \lambda \cdot L_s \tag{15}$$

4 Experiments

We evaluated DiffSQL on two public datasets, KITTI and Make3D, and use widely adopted metrics [Eigen and Fergus, 2015] to quantify performance. The model's generalization ability is assessed through zero-shot evaluation.

4.1 Datasets and Experimental Protocol

KITTI. The KITTI stereo dataset [Geiger *et al.*, 2013] comprises 61 driving scenarios acquired by synchronized stereo cameras and LiDAR (resolution 1242×375). Adopting Eigen's benchmark split [Eigen *et al.*, 2014], we utilize 39,810 monocular triplets for training and 4,424 for validation. The test set contains raw LiDAR measurements (697 frames) and sparsity-corrected ground truth [Uhrig *et al.*, 2017] (652 frames).

Make3D. The Make3D dataset [Saxena *et al.*, 2008] validates DiffSQL's cross-dataset generalization capability by employing zero-shot evaluation with KITTI-pre-trained weights.

Method	Type	AbsRel↓	SqRel↓	RMSE↓	log_{10}↓
Monodepth	S	0.544	10.94	11.760	0.193
Zhou	M	0.383	5.321	10.470	0.478
DDVO	M	0.387	4.720	8.090	0.204
Monodepth2	M	0.322	3.589	7.417	0.163
SQLdepth	M	0.314	3.374	7.285	0.161
DiffSQL(Ours)	M	**0.310**	**3.013**	**7.019**	**0.159**

Table 3: Make3D results

4.2 Implementation Details

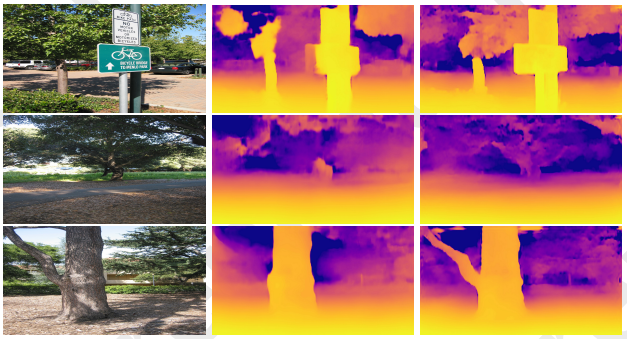
The results in Table 1 show that DiffSQL outperforms all existing self-supervised methods, including those that use stereo pairs or multiple frames. A comparative analysis with Monodepth2 and SQLDepth demonstrates that DiffSQL excels at preserving details of distant and small objects, as shown in Fig. 1 and Fig. 6. Furthermore, with improved ground truth from KITTI [Uhrig *et al.*, 2017], DiffSQL outperforms SQLdepth [Wang *et al.*, 2024] on all metrics in Table 2.

4.3 Zero-Shot Generalization on Make3D

For zero-shot evaluation, we used KITTI-pretrained weights to test on the Make3D dataset [Saxena *et al.*, 2008]. Following [Godard *et al.*, 2017], we tested on a center-cropped image with a 2:1 aspect ratio. As shown in Table 3 and Fig. 7, DiffSQL generates sharper depth maps with more accurate scene details, demonstrating strong zero-shot generalization.

4.4 Ablation Study

This section conducts ablation studies to examine the effects of different modules on DiffSQL, such as diffusion-based feature fusion, various layers of the SD model, and the dynamic query mechanism.



(b) RGB (b) SQLdepth (c) DiffSQL(Ours)

Figure 7: Qualitative Make3D results (Zero-shot).

Ablation	AbsRel↓	SqRel↓	RMSE↓
ResNet18	0.112	0.878	4.602
ResNet50	0.103	0.758	4.499
ResNet18 + StableDiffusion	0.099	0.704	4.436
ResNet50 + StableDiffusion	**0.096**	**0.698**	**4.333**

Table 4: The impact of feature fusion modules based on different convolutional network architectures on depth maps.

Effect of Diffusion-based Feature Fusion. As shown in Table 4, we investigate the effect of combining Stable Diffusion (SD) with convolutional networks on model performance. The results in 4 reveal that the "ResNet50 + Stable Diffusion" configuration surpasses the base ResNet50 model in all evaluation metrics (AbsRel, SqRel, RMSE). This improvement highlights the ability of SD to capture fine-grained scene details, especially distant and small objects. By fusing the features extracted by ResNet and stable diffusion, we achieve superior performance with minimal error. This showcases that SD's strong feature extraction and image refinement capabilities and validating the potential of generative models in depth estimation tasks.

Effect of Different SD Model Layers. As shown in Table 5, we investigate the impact of various SD model layers on depth estimation. Three settings are compared: Mid Block (0), Up Block (1), and Up Block (0). The extracted features from various SD layers, utilized as coarse-grained query objects, markedly influence the accuracy of depth estimation. Notably, the Up Block (0) configuration excels in capturing fine details and enhancing depth accuracy, underscoring the importance of the SD model layer selection in depth estimation.

Effect of Dynamic Query Layer. As shown in Table 6, we compare the effects of dynamic query mechanisms and fixed queries on DiffSQL. The dynamic query mechanism adapts queries to the scene features, enhancing depth-interval flexibility. Experimental results demonstrate that this mechanism allows the model to focus on key regions within coarse-grained query objects, improving the capture of distant and small objects. This adaptive strategy improves depth map accuracy and improves performance in complex scenes.

Ablation	AbsRel↓	SqRel↓	RMSE↓
Mid Block (0)	0.100	0.732	4.565
Up Block (1)	0.098	0.712	4.547
Up Block (0)	**0.096**	**0.698**	**4.333**

Table 5: The influence of features extracted from different layers of the SD model on depth maps.

Ablation	AbsRel↓	SqRel↓	RMSE↓
No queries	0.105	0.788	4.623
Fixed queries	0.101	0.748	4.512
Dynamic queries	**0.096**	**0.698**	**4.333**

Table 6: The effect of dynamic query layer on depth maps.

5 Conclusion

In this study, we propose DiffSQL, a framework that leverages the generative priors of Stable Diffusion to augment geometric feature extraction in convolutional networks while integrating an adaptive query modulation mechanism. The synergistic architecture demonstrates superior performance in capturing geometric details of distant and small-scale objects compared to existing approaches. Systematic benchmarking on standard datasets reveals marked improvements in both depth accuracy and cross-domain generalization. This work illuminates the transformative potential of generative models such as Stable Diffusion in addressing the inherent challenges of monocular depth estimation.

Acknowledgements

This work is supported by National Key R&D Program of China (No. 2024YFC3014300) and by the Natural Science Foundation of China under Grants No. 62372378 and 72225011.

References

[Achtelik *et al.*, 2009] Markus Achtelik, Abraham Bachrach, Ruijie He, Samuel Prentice, and Nicholas Roy. Stereo vision and laser odometry for autonomous helicopters in gps-denied indoor environments. In *Unmanned Systems Technology XI*, volume 7332, pages 336–345. SPIE, 2009.

[Baranchuk *et al.*, 2021] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[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*, 2024.

[Bhat *et al.*, 2021] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference*

on computer vision and pattern recognition, pages 4009–4018, 2021.

[Chen *et al.*, 2023] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19830–19843, 2023.

[Chen *et al.*, 2024] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[Diaz and Marathe, 2019] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4738–4747, 2019.

[Eigen and Fergus, 2015] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[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.

[Fu *et al.*, 2018] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018.

[Garg *et al.*, 2016] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 740–756. Springer, 2016.

[Garg *et al.*, 2020] Divyansh Garg, Yan Wang, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Wasserstein distances for stereo disparity estimation. *Advances in Neural Information Processing Systems*, 33:22517–22529, 2020.

[Geiger *et al.*, 2013] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[Godard *et al.*, 2019a] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019.

[Godard *et al.*, 2019b] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019.

[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.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Huynh *et al.*, 2020] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä. Guiding monocular depth estimation using depth-attention volume. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 581–597. Springer, 2020.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[Koh *et al.*, 2024] Junyoung Koh, Sanghyun Park, and Joy Song. Improving text generation on images with synthetic captions. In *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 644–649. IEEE, 2024.

[Liu *et al.*, 2023] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. Va-depthnet: A variational approach to single image depth prediction. *arXiv preprint arXiv:2302.06556*, 2023.

[Lyu *et al.*, 2021] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2294–2301, 2021.

[Namekata *et al.*, 2023] Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[Newcombe *et al.*, 2011] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE*

international symposium on mixed and augmented reality, pages 127–136. Ieee, 2011.

[Rombach *et al.*, 2021] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[Saharia *et al.*, 2022] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[Saxena *et al.*, 2008] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008.

[Shao *et al.*, 2023] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Nddepth: Normal-distance assisted monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7931–7940, 2023.

[Shao *et al.*, 2024] Shuwei Shao, Zhongcai Pei, Weihai Chen, Dingchi Sun, Peter CY Chen, and Zhengguo Li. Monodiffusion: self-supervised monocular depth estimation using diffusion model. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[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.

[Wang *et al.*, 2024] Youhong Wang, Yunji Liang, Hao Xu, Shaohui Jiao, and Hongkai Yu. Sqldepth: Generalizable self-supervised fine-structured monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5713–5721, 2024.

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1164–1174, 2021.

[Wolleb *et al.*, 2022] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022.

[Yan *et al.*, 2021] Jiaxing Yan, Hong Zhao, Penghui Bu, and YuSheng Jin. Channel-wise attention-based network for self-supervised monocular depth estimation. In *2021 International Conference on 3D vision (3DV)*, pages 464–473. IEEE, 2021.

[Yuan *et al.*, 2022] W Yuan, X Gu, Z Dai, S Zhu, and P Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. arxiv 2022. *arXiv preprint arXiv:2203.01502*, 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.

[Zhou *et al.*, 2017] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017.