2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Self-Supervised Monocular Depth Estimation Based on Dual-Branch DepthNet and Multi-Attention Fusion

Yuanyuan Wang[1], Dianxi Shi[2(✉)], Junze Zhang[1], Luoxi Jing[3],
Xueqi Li[1], and Xucan Chen[1]

[1] Academy of Military Science, Beijing 100850, China
[2] Intelligent Game and Decision Lab (IGDL), Beijing 100091, China
[3] School of Computer Science, Peking University, Beijing, 100871, China
dxshi@nudt.edu.cn

Abstract. Monocular depth estimation infers 3D geometric structures of scenes from a single RGB image, offering significant applications in autonomous driving, robot navigation, and other fields. While current self-supervised learning methods avoid dependency on ground truth depth data, they still exhibit notable limitations in complex scenarios: traditional encoder-decoder architectures inevitably lose high-frequency detail features when acquiring global context through continuous downsampling, resulting in blurred edges and texture distortion in depth maps. To alleviate these issues, we propose a novel approach named HyperDetailNet, which significantly enhances depth estimation detail preservation. Specifically, our method contains two key components: 1) A dual-branch detail-global feature extraction network, where the detail branch adopts an enlarge-then-reduce strategy to preserve high-frequency texture information, while the global branch extracts overall structural information of the scene. 2) To effectively fuse features from both branches, we designed a multi-attention fusion module that combines spatial attention, channel attention, and sliding window self-attention mechanisms to enhance model perception of detailed regions. Experimental results demonstrate that HyperDetailNet achieves excellent performance on both KITTI and Make3D datasets, with significant improvements in depth estimation for edge and texture-rich areas. Additionally, ablation experiments verify the effectiveness of the dual-branch detail-global feature extraction DepthNet and multi-attention fusion module.

Keywords: Monocular Depth Estimation • Self-Supervised • Dual-Branch Network

1 Introduction

Monocular depth estimation, as one of the crucial tasks in computer vision, has demonstrated immense potential in various applications including autonomous driving [1], robot navigation [2], and augmented reality [3]. In these applications, depth information is essential for understanding the environment, making decisions, and conducting accurate spatial positioning. Particularly in autonomous driving, precise depth estimation helps vehicles comprehend the 3D structure of their surroundings, enabling safe and swift decision-making. However, despite significant advancements in depth estimation technology, accurately extracting fine-grained depth information from monocular images, especially in texture-rich regions, remains a formidable challenge.

Traditional depth estimation methods [4-6] typically rely on multi-view or stereo vision techniques, obtaining depth information by matching image pairs from multiple viewpoints. The advantage of these methods lies in their ability to utilize disparity information from multiple perspectives to acquire relatively accurate depth maps. However, stereo vision methods [7,8] require additional camera hardware support and showcase suboptimal performance in smooth regions or scenes with significant view changes. With the proliferation of deep learning technology, researchers have proposed supervised monocular depth estimation methods [9-12] that can directly predict depth information from a single image through training on large-scale annotated data. Although supervised learning methods can improve accuracy through extensive labeled data, acquiring precise depth data is extremely difficult and costly. Traditional depth sensors (such as LiDAR) struggle to obtain comprehensive or high-quality depth information in certain scenarios, while manual annotation of depth data is not only time-and labor-intensive but also challenging in covering various complex environments and lighting conditions. Therefore, practical applications of monocular depth estimation often face the issue of insufficient labeled data, thus casting shadows on the widespread application of supervised methods.

Self-supervised monocular depth estimation [13-17] has gradually become a popular research in recent years to overcome the dependency on labeled data in supervised learning. Self-supervised methods utilize geometric and photometric consistency between images for training, eliminating the need for ground truth depth data. These methods can improve depth estimation accuracy by learning internal structural relationships within images without labeled data. However, many existing self-supervised methods typically extract multi-scale feature maps through consecutive downsampling operations [13,14,18-20], which, while capturing global information, result in progressive reduction of feature map resolution, making the extraction of detailed information insufficient. This gradually decreasing resolution leads to the loss of high-frequency textures and local details, affecting the detail presentation of depth maps, especially in complex or high-texture regions.

To address these issues, we propose a novel self-supervised monocular depth estimation method named HyperDetailNet. This method effectively enhances the detail presentation of depth maps through a dual-branch detail-global feature extraction network that introduces a detail branch with hierarchical upsampling high-frequency feature perception. Specifically, we incorporate a detail branch in the depth network dedicated to extracting high-frequency texture information, with the global branch to ensure global scene information without losing important local details. To further integrate these two types of features, we propose a multi-attention fusion module to enhance detail extraction and global perception capabilities. In processing the feature maps extracted by the detail branch, we introduce channel and spatial attention mechanisms to ensure the model can selectively enhance features in key regions. Meanwhile, sliding window self-attention computation is applied to the feature maps extracted by the global branch, further improving the ability to capture long-range feature dependencies. Through this multi-attention fusion approach, the accuracy and robustness of depth estimation are enhanced. Our contributions mainly include the following aspects:

- We design a dual-branch detail-global feature extraction network where the detail branch adopts an enlarge-then-reduce strategy, focusing on extracting high-frequency texture information; the global branch is based on the traditional U-Net structure, responsible for extracting the overall semantic information of the scene. This dual-branch architecture can simultaneously attend to local details and global structures, thereby improving the accuracy and detail presentation of depth estimation.
- We propose a multi-attention fusion module that combines spatial attention, channel attention, and sliding window self-attention mechanisms to effectively merge features from the detail and global branches. Through this multi-level attention mechanism, our model can adaptively focus on important regions and channels in the image, enhancing detail perception and global understanding capabilities.
- Experimental results demonstrate that HyperDetailNet achieves superior performance compared to existing methods on both KITTI [21] and Make3D [22] datasets. Especially in zero-shot testing on the Make3D dataset, our method exhibits excellent generalization capability. Ablation experiments further verify the effectiveness and necessity of the dual-branch detail-global feature extraction network and multi-attention fusion module.

2 Related Work

2.1 Supervised Depth Estimation

Early depth estimation methods largely relied on supervised learning, using ground truth depth data provided by devices such as LiDAR or stereo vision systems as labels for training. These methods typically employ regression or classification approaches to directly predict the depth of each pixel from images. In 2014, Eigen et al. [9] first proposed a supervised monocular depth estimation framework based on convolutional neural networks (CNN), adopting a multi-scale network structure: a global network extracts overall scene information, a local network refines details, and depth maps are generated by fusing multi-scale features. Subsequently, Laina et al. [10] introduced ResNet [23] to enhance model depth and optimization capabilities, and proposed sub-pixel convolution upsampling to improve resolution. To further address the challenges of wide depth ranges and optimization difficulties in regression tasks, researchers redefined depth estimation as a joint classification-regression task. For instance, AdaBins proposed by Bhat et al. [11] dynamically divides depth ranges into adaptive bins, predicting depth distribution probabilities through classification and then regressing weighted bin center values, significantly improving the continuity and accuracy of depth estimation. In recent years, architectures such as Transformer [24] and Swin-Transformer [25] have been introduced to enhance long-range feature modeling capabilities through global attention mechanisms (e.g., NeWCRFs [12]). However, supervised methods rely on large-scale annotated data, and the acquisition of depth ground truth is costly, while sensors face issues of sparsity and noise in complex scenes, limiting the generalization capabilities of models.

2.2 Self-supervised Monocular Depth Estimation

Stereo Image Training. Garg et al. [7] first proposed a photometric consistency constraint based on stereo images, reconstructing the image from another viewpoint by predicting disparity maps and using reconstruction error as a supervision signal. Building on this, Godard et al. [8] proposed left-right disparity consistency constraints and introduced the Structural Similarity Index (SSIM) as part of the reconstruction loss, making the training process more stable and robust, thereby significantly improving the accuracy of depth estimation. The stereo image training approach has the advantage of directly obtaining absolute scale information while effectively utilizing fixed stereo geometric constraints for depth reconstruction in static scenes. However, its limitations are also quite evident. On one hand, this method requires acquiring strictly calibrated stereo image pairs, and in practical applications, the installation and calibration costs of stereo sensors are relatively high. On the other hand, due to the interference of occlusion, reflection, and dynamic object movement in stereo imaging, it often struggles to recover detailed information and boundary structures when processing complex scenes, resulting in blurring in the detailed parts of depth maps. Additionally, self-supervised methods based on stereo images also have certain deficiencies in handling issues such as lighting changes, scene dynamics, and smooth regions, limiting their widespread application in large-scale self-supervised learning.

Monocular Video Training. To overcome the limitations of the stereo training paradigm, researchers began to turn to self-supervised training using monocular video sequences. The advantage of monocular video training lies in the more convenient data acquisition, as almost all mobile devices can capture videos, making massive unlabeled data possible. At the same time, utilizing motion information between consecutive frames provides more geometric constraints and temporal consistency, which to some extent alleviates issues such as occlusion, dynamic objects, and lighting changes. Zhou et al. [13] jointly trained a DepthNet and a PoseNet in SfMLearner, using photometric loss as a supervision signal. DepthNet is used to predict the depth map from a single frame, while PoseNet predicts the relative pose between adjacent frames. The network parameters are optimized by synthesizing reference frame images and comparing them with source frame images to calculate photometric loss, thereby achieving depth estimation. Subsequently, Godard et al. [14] further improved this framework in their Monodepth2, proposing a more robust monocular video training strategy that employs an automatic masking mechanism to filter out unreliable regions caused by dynamic objects or stationary cameras, while combining multi-scale features and various reconstruction losses in the reconstruction loss, making the training process more efficient and stable. Besides the above methods, some researchers have introduced additional prior information or multi-task learning frameworks based on monocular video training. For example, Mahjourian et al. [15] proposed using 3D geometric consistency to further constrain depth prediction, ensuring that point clouds reconstructed from adjacent frames maintain consistency in 3D space, thereby improving the global consistency and detail recovery capability of depth maps. Feng et al. [16] addressed the instability issue of depth estimation in dynamic scenes by proposing a dynamic depth estimation method, which further improves the adaptability and robustness of monocular video training in practical scenarios through the introduction of motion separation mechanisms and dynamic target detection. Meanwhile, Watson et al. [17] achieved a breakthrough in utilizing multi-frame information with their ManyDepth, which enables the model to better capture detailed information in monocular video sequences and maintain high estimation accuracy in complex scenes through multi-view fusion and fine-grained feature alignment strategies. These methods have solved, to varying degrees, the problems caused by dynamic objects, occlusion, and lighting changes in monocular video training, while also demonstrating the feasibility and advantages of utilizing geometric and photometric information between consecutive video frames for depth estimation.

2.3 Self-supervised Monocular Depth Estimation Methods for Edge Problems

To address edge detail issues, researchers have proposed various improvement methods. Yang et al. [26] proposed an edge-aware depth estimation network, explicitly guiding the network to focus on object contours by introducing an edge detection branch. Ramamonjisoa et al. [27] designed SharpNet, which simultaneously predicts depth, surface normal vectors, and edges, improving depth accuracy in edge regions through multi-task learning. In recent years, attention mechanisms have been widely applied to enhance the detail preservation capability of depth estimation. Xu et al. [28] proposed a multi-scale attention network, effectively enhancing the representation of edges and details by applying attention modules at different feature scales. Johnston et al. [29] proposed self-supervised attention-guided depth estimation, adaptively focusing on key regions in images through spatial and channel attention mechanisms, improving detail preservation capability. Zhang et al. [30] introduced an attention network based on feature pyramids, balancing the extraction of global and local information by applying attention mechanisms on multi-scale features. Although the above methods have improved edge detail issues partly, it is worth noting that most of them still adopt a layer-by-layer downsampling approach for feature extraction, a strategy that leads to continuous reduction in feature map resolution and gradual loss of high-frequency information. Particularly in deep networks, the resolution of feature maps may decrease to $1/32$ of the original input or even lower, severely limiting the ability of the model to express details and ultimately resulting in insufficient accuracy of predicted depth maps in object edges and texture-rich regions. Moreover, these methods typically employ a single-branch architecture, making it difficult to simultaneously consider global scene understanding and local detail extraction. The HyperDetailNet proposed in this paper more effectively addresses this issue through dual-branch detail and global branch structures, as well as a multi-attention fusion module, significantly improving the depth estimation accuracy in edge and high-frequency texture regions while maintaining global consistency.

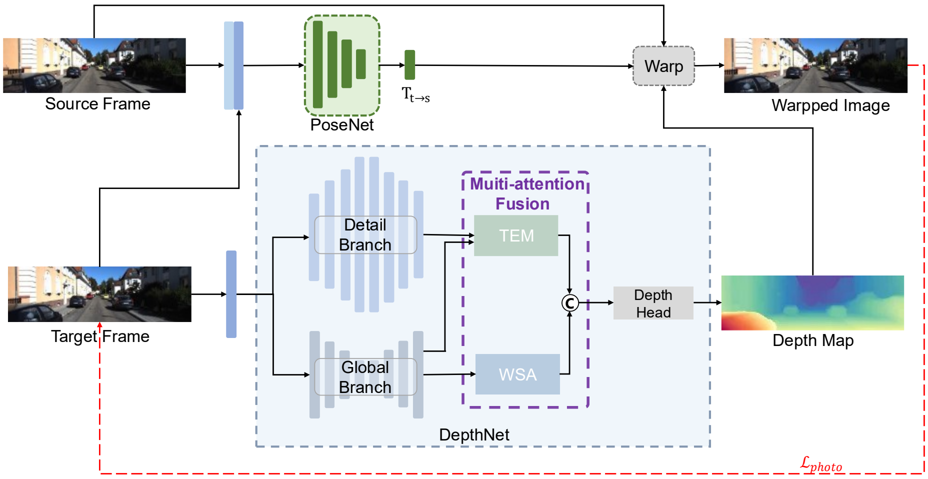

Fig. 1. Overview of our HyperDetailNet architecture. HyperDetailNet comprises two core components: a DepthNet for predicting depth maps and a PoseNet for estimating relative camera poses between adjacent monocular frames. In DepthNet, a dual-branch detail-global dual-branch feature extraction architecture is designed to capture both detail and global features from input images. A Multi-attention Fusion (MAF) module is further incorporated to adaptively enhance and integrate these complementary feature representations through channel-wise, spatial attention and window self-attention mechanisms. The MAF module consists of two parts, Texture Enhanced Module (TEM) and Window Self-Attention (WSA) module.

3 Method

This section details our proposed HyperDetailNet method, which significantly enhances detail representation in self-supervised monocular depth estimation through dual-branch detail-global feature extraction networks and multi-attention fusion mechanisms. As shown in **Fig. 1**, our network consists of three main components: a dual-branch detail-global feature extraction DepthNet, a multi-attention fusion module, and a PoseNet. The following will explain the design concepts and specific implementation of each component.

3.1 Dual-Branch Detail-Global Feature Extraction DepthNet

Traditional depth estimation networks [14,31] typically adopt a single encoder-decoder structure, extracting multi-scale features through consecutive downsampling operations. While this approach effectively captures global scene information, it leads to the loss of high-frequency details, especially in object edges and texture-rich regions. To address this issue, we propose a dual-branch structure. As shown in **Fig. 2**, the dual-branch detail-global feature extraction DepthNet includes detail and global branches, focusing on high-frequency texture information and global scene understanding respectively. These branches share the same input image but employ different feature

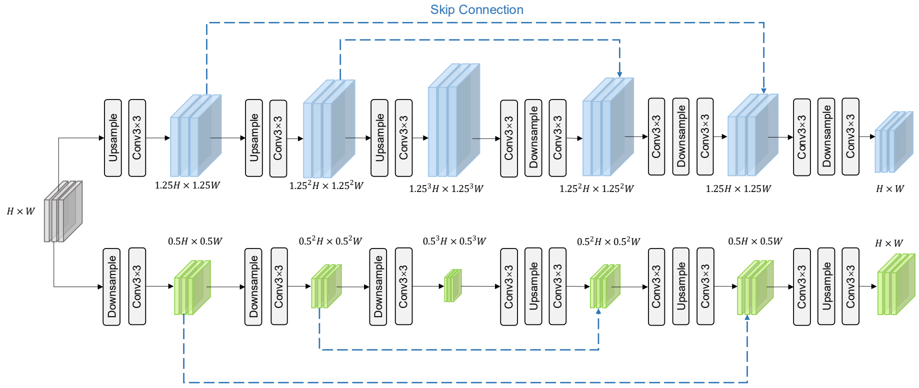

Fig. 2. Structures of Dual-Branch Detail-Global Feature Extraction DepthNet. The upper part is the detail branch, and the lower part is the global branch.

extraction strategies. Given the current input image (size 192×640 in training), we first apply a 3×3 convolution with $stride = 2$ to downsample the input image to $1/2$ of the original resolution, producing feature map $F \in \mathbb{R}^{C \times H \times W}$. This initial downsampling operation aims to reduce computational complexity while retaining sufficient spatial information. Feature map F is then input in dual-branch to both detail and global branches for further processing.

Detail Branch. The detail branch adopts an enlarge-then-reduce strategy, similar to U-Net [32] architecture. Its core design concept is to expand the spatial resolution of feature maps in the early stages of feature extraction to preserve high-frequency texture information. In the encoding stage, feature map size is gradually expanded through a series of upsampling-convolution modules. Each upsampling operation is followed by a 3×3 convolution layer to refine local details in the feature map. This design enables the network to retain more high-frequency texture information during feature extraction, providing richer detail representations for subsequent depth estimation. Given input feature F, the feature map at layer i is represented as:

$$F_i = Conv_{3\times3}\big(Upsample(F_{i-1})\big) \tag{1}$$

where $Upsample(\cdot)$ represents the upsampling operation implemented through bilinear interpolation, expanding the spatial dimensions of the feature map by a factor of 1.25. $Conv_{3\times3}(\cdot)$ represents a 3×3 convolution with $stride = 1$. The decoding stage restores the feature map size to the original input size through consecutive downsampling-convolution modules, obtaining detail features at different scales.

Global Branch. The global branch adopts a U-Net architecture based on ResNet blocks, following the design concept of [14], focusing on extracting global semantic information and scene structure from images. Given input feature F, the global branch first gradually reduces the spatial dimensions of feature maps through a series of downsampling operations while increasing the number of channels to expand the receptive field. In the decoding stage, the global branch gradually restores the spatial resolution of feature maps through upsampling operations and fuses corresponding layer features from the encoding stage through skip connections to preserve more spatial details. This encoder-decoder structure enables the global branch to effectively understand the overall structure and semantic information of the scene, providing global constraints for depth estimation.

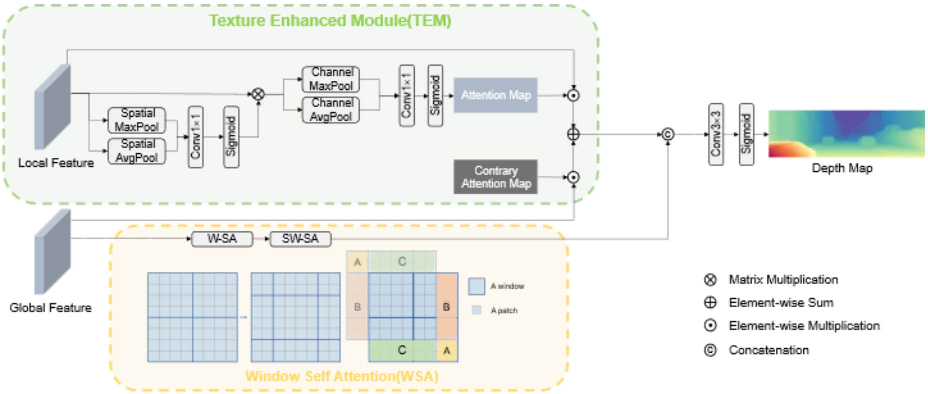

Fig. 3. Structures of Multi-attention Fusion Module. MAF incorporates a texture enhanced module for enhancing detail features and a window self-attention component for enhancing global features.

3.2 Multi-attention Fusion Module

To efficiently fuse features extracted from the detail and global branches, we propose a Multi-attention Fusion (MAF) module. As shown in **Fig. 3**, this module includes a texture enhanced module and window self-attention, combining channel attention, spatial attention, and sliding window self-attention mechanisms to improve both detail feature extraction and global perception capabilities, thereby better capturing and integrating local and global information in images.

Texture Enhanced Module. In processing feature maps extracted by the detail branch, we introduce channel attention and spatial attention mechanisms, derived from Convolutional Block Attention Module (CBAM) [33]. This module receives features of scale $H \times W$ from the detail branch decoding stage. It first processes the feature map through Spatial MaxPool and Spatial AvgPool operations, capturing salient features and overall distribution from the spatial dimension. The pooled features generate a spatial attention map through shared convolution layers and Sigmoid activation, guiding the network to focus on important regions in the image. After element-wise multiplication of the spatial attention map with the initial detail features, we capture important difference between channels using Channel MaxPool and Channel AvgPool, ensuring the model adaptively focuses on specific regions and channels. The pooled features generate the final attention map M_{att} through convolution layers and Sigmoid activation. Through these operations, this attention map M_{att} has high weights in image detail texture regions, as these have high feature values in detail features. Similarly, it has lower weights in smooth regions. Thus, we obtain a reverse attention map by inverting the attention map, emphasizing regions overlooked by the original attention mechanism. The texture-enhanced detail feature representation is obtained through the following formula:

$$F_{detail}^{enhanced} = F_{detail} \odot M_{att} \oplus F_{global} \odot (1 - M_{att}) \tag{2}$$

Window Self Attention. To further enhance global feature extraction capabilities, we introduce window self-attention mechanisms in processing the global branch feature maps, referencing [34] and adopting Swin Transformer layers (STL) [25] to preserve contextual information while alleviating GPU memory issues. We divide the global features F_{global} output from the global branch into fixed-size windows and calculate relationships between features within each window through window self-attention (W-SA). Then, through sliding window self-attention (SW-SA), we expand the feature propagation range, enabling information exchange between windows. The enhanced global features can be expressed as:

$$F_{global}^{enhanced} = SW\text{-}SA\left(W\text{-}SA\left(F_{global}\right)\right) \tag{3}$$

Through this method, the model can capture global information across a wider range while avoiding the high computational complexity of traditional self-attention calculations. Finally, the detail features $F_{detail}^{enhanced}$ processed by the texture enhancement module and the global features $F_{global}^{enhanced}$ processed by window self-attention are concatenated along the channel dimension and processed through a 3×3 convolution layer and Sigmoid activation function to generate the final depth map:

$$D = \sigma\left(Conv_{3\times3}\left(Concat\left(F_{detail}^{enhanced}, F_{global}^{enhanced}\right)\right)\right) \tag{4}$$

where σ represents the Sigmoid activation function, $Conv_{3\times3}$ represents the convolution operation, and $Concat$ represents the feature concatenation operation along the channel dimension.

3.3 Self-supervised Learning

The core concept of self-supervised learning is utilizing geometric relationships between consecutive video frames for training. Specifically, given a single input image I_t, the DepthNet predicts its corresponding depth map D_t. The PoseNet then takes two temporally adjacent images (the current frame I_t and an adjacent frame I_s) as input and predicts the relative pose $T_{t\to s}$ between the target image I_t and the adjacent image I_s. Finally, view synthesis is performed using the adjacent frame I_s, pose transformation matrix $T_{t\to s}$, and depth map D_t to obtain a reconstructed image $\hat{I}_t$ of the target image I_t. The DepthNet and PoseNet are jointly optimized by minimizing the image reconstruction loss $\mathcal{L}_r$ between the target image I_t and the reconstructed image $\hat{I}_t$, as well as an edge-aware smoothness loss $\mathcal{L}_{smooth}$ constraining the depth map.

Image Reconstruction Loss. The photometric reprojection error composed of L1 loss and Structural Similarity (SSIM) [35] is shown as follows:

$$\mathcal{L}_p\left(\hat{I}_t, I_t\right) = \alpha \frac{1 - SSIM(\hat{I}_t, I_t)}{2} + (1 - \alpha)\left\|\hat{I}_t - I_t\right\| \tag{5}$$

where α is empirically set to 0.85 [13]. In real-world scenarios, stationary cameras or moving objects break the assumption of a moving camera and static scene, damaging the training process of self-supervised depth estimation. To address this issue, we use the automatic masking strategy from Monodepth2 to filter out stationary pixels and low-texture regions that maintain the same appearance between two frames in a sequence. The masking coefficient μ is calculated in the formula below, where [] is the Iverson bracket.

$$\mu = \left[\min_s \mathcal{L}_p(I_s, I_t) > \min_s \mathcal{L}_p(\hat{I}_t, I_t)\right] \tag{6}$$

Therefore, the image reconstruction loss is defined as:

$$\mathcal{L}_r(\hat{I}_t, I_t) = \mu \cdot \mathcal{L}_p(I_s, I_t) \tag{7}$$

Edge-aware Smoothness Loss. Inspired by previous works [14,36,37], to regularize depths in textureless regions, we use an edge-aware smoothness loss as regularization, as follows:

$$\mathcal{L}_{smooth} = |\partial_x \mathrm{d}_t^*| e^{-|\partial_x I_t|} + |\partial_x \mathrm{d}_t^*| e^{-|\partial_y I_t|} \tag{8}$$

where $d_t^* = d_t / \hat{d}_t$ represents the mean-normalized inverse depth. Therefore, the total loss can be represented as:

$$\mathcal{L} = \frac{1}{S} \sum_i^S (\mathcal{L}_r + \lambda \mathcal{L}_{smooth}) \tag{9}$$

where S is the number of scales, and λ is the weight of the smoothness regularization term. λ is set to 0.001 as in [14].

4 Experiments

We evaluated the HyperDetailNet model on KITTI and Make3D datasets with seven standard metrics. We also studied the generalization ability of the model on unseen datasets through zero-shot evaluation experiments.

4.1 Datasets and Evaluation Metrics

KITTI. KITTI [21] is a standard benchmark for depth estimation in autonomous driving scenarios. It contains image sequences from urban, rural, and highway scenes captured from moving vehicles. Following Eigen split [38], we used images from 29 scenes (totally 39,810 images) for training and 697 images from different scenes for testing. During evaluation, predicted depths are usually limited to the [0,80]m range.

Table 1. Comparison of HyperDetailNet with some recent representative methods on the KITTI benchmark using the Eigen split. All input images are resized to 640 × 192 unless otherwise specified. "M": KITTI monocular videos, "M†": without pre-training on ImageNet [39]. Best results are in **bold**. Our method performs better when pre-trained weights are loaded.

Method	Data	Depth Error(↓)				Depth Accuracy(↑)		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta <$ 1.25	$\delta <$ 1.25^2	$\delta <$ 1.25^3
Monodepth2 [14]	M	0.115	0.903	4.872	0.193	**0.877**	0.959	0.981
Struct2Depth [41]	M	0.141	1.026	5.291	0.215	0.816	0.945	0.979
SGDepth [31]	M	0.117	0.907	4.844	0.196	0.875	0.958	0.980
Li et al. [42]	M	0.130	0.950	5.138	0.209	0.843	0.948	0.978
Zhang et al. [43]	M	0.173	1.153	4.979	0.249	0.752	0.916	0.968
Heydrich et al. [44]	M	0.145	1.154	5.775	0.238	0.792	0.924	0.969
MonoDA [45]	M	0.126	1.035	5.105	0.203	0.857	0.955	0.979
Dynamo-Depth [46]	M	0.120	0.864	4.850	0.195	0.858	0.956	0.982
ColorDepth [47]	M	**0.114**	0.846	4.839	0.196	0.853	0.953	0.982
Ours	M	0.118	**0.839**	**4.748**	**0.192**	0.871	**0.960**	**0.982**
Monodepth2 [14]	M†	0.132	1.044	5.142	0.210	0.845	0.948	0.977
R-MSFM3 [18]	M†	**0.128**	0.965	5.019	**0.207**	**0.853**	**0.951**	0.977
Ours	M†	0.132	**0.935**	**5.007**	0.208	0.841	0.949	**0.978**

Table 2. Comparison of the proposed HyperDetailNet to some other methods on the Make3D dataset. All models are trained on KITTI with an image resolution of 640 × 192. Best results are in **bold**.

Method	Data	Abs Rel	Sq Rel	RMSE	RMSE log
Zhou et al. [48]	M	0.383	5.321	10.470	0.478
DDVO [49]	M	0.387	4.720	8.090	0.204
Monodepth2 [14]	M	0.322	3.589	7.417	0.163
R-MSFM6 [18]	M	0.334	3.285	7.212	0.169
CADepth [19]	M	0.319	3.564	7.152	0.158
Xu et al. [50]	M	0.466	7.052	9.568	0.179
ColorDepth [47]	M	0.338	3.427	7.646	0.168
Ours	M	**0.314**	**3.111**	**7.023**	**0.161**

Make3D. Make3D [22] dataset contains images of outdoor static scenes with corresponding depth maps. It features more diverse scenes than KITTI, including buildings, trees, and open spaces. We used our model trained on KITTI to directly evaluate standard on Make3D test set of 134 images to assess generalization capability.

Evaluation Metrics. To comprehensively evaluate depth estimation performance, we adopted seven standard metrics: (1) Absolute Relative Error (AbsRel): Reflects the average relative difference between predicted and actual values. (2) Squared Relative Error (SqRel): More sensitive to larger depth errors by amplifying prediction deviations in distant areas. (3) Root Mean Square Error (RMSE): Calculates the global root mean square difference between predicted and actual values. (4) Log Root Mean Square Error (RMSE log): Calculates errors in logarithmic space to balance near and far-field error contributions. (5) Threshold Accuracy ($\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$): Measures the percentage of pixels where the ratio between prediction and ground truth falls within specified thresholds. These metrics evaluate model performance across error distribution, extreme value sensitivity, scale robustness, and confidence.

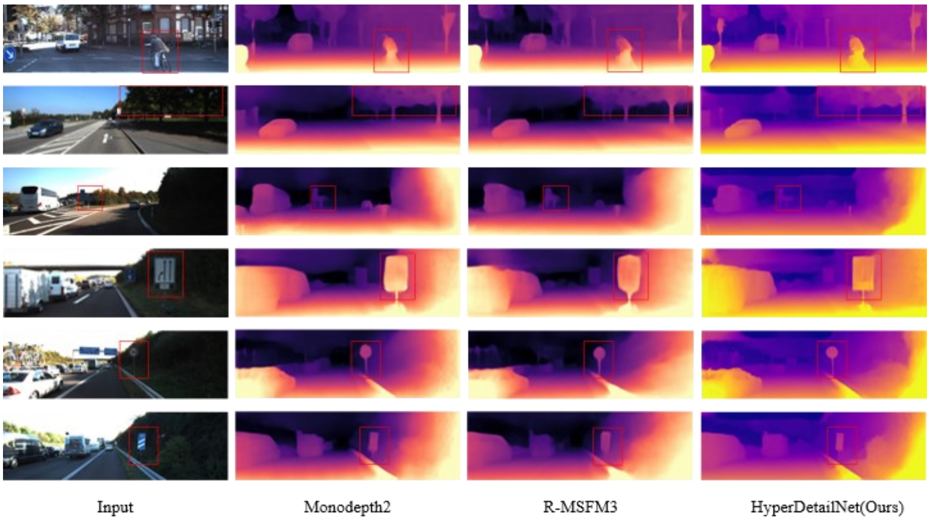

Fig. 4. Qualitative results on KITTI.

4.2 Implementation Details

Training Configuration. Our implementation is based on PyTorch framework, trained for 20 epochs on KITTI. We used the Adam optimizer with a batch size of 8 on multiple RTX 3090 GPUs. Input images were resized to a standard resolution 640 × 192. The learning rate was set to 10^{-4} initially for the first 15 epochs, then reduced to 10^{-5} to ensure convergence.

Data Augmentation. To improve model robustness, we followed previous methods [14,18,40], applying augmentation during preprocessing. Specifically, we randomly performed horizontal flipping, brightness adjustment (± 0.2), saturation adjustment (± 0.2), contrast adjustment (± 0.2), and hue jittering (± 0.1) with a 50% probability.

4.3 Results

Results on KITTI Dataset. As shown in **Table 1**, the comparison between HyperDetailNet and current mainstream self-supervised monocular depth estimation methods on KITTI proves the superiority of our method. All methods used the same input resolution (640 × 192) and evaluation metrics for fair comparison. Our experiments were divided into two groups: without pre-trained weights (M) and with ImageNet pre-trained weights (M†). With pre-trained weights, our HyperDetailNet achieved comparable performance to existing methods. In particular, our method improves the RMSE score by about 12.4 % and Squared Relative Error by 6.4% over the classical method Monodepth2, indicating our method can generate more accurate depth boundaries and structural details. Without ImageNet pre-trained weights, though overall performance declined, our method still showed competitiveness compared to others. On SqRel and RMSE metrics, our method achieved 0.935 and 5.007 respectively, showing competitiveness compared to R-MSFM3. **Fig. 4** shows qualitative results, clearly demonstrating the advantages of HyperDetailNet in preserving object edges and texture details. Especially in high-frequency detail areas like building outlines, road signs, and vegetation, our method generates depth maps with clearer boundaries and richer details.

Table 3. Ablation study on model architectures. All the models are trained and tested on KITTI with the input size 640 × 192.

Architecture	Depth Error(↓)				Depth Accuracy(↑)		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta <$ 1.25	$\delta <$ 1.25^2	$\delta <$ 1.25^3
Full model	**0.118**	**0.839**	**4.748**	**0.192**	**0.871**	**0.960**	**0.982**
w/o Dual-Branch DepthNet	0.120	0.924	4.917	0.197	0.865	0.958	0.980
w/o Multi-attention Fusion	0.120	0.901	4.844	0.194	0.869	0.957	0.981

Results on Make3D. To evaluate generalization capability, we directly applied our KITTI-trained model to the Make3D test set without fine-tuning. We trained our model on 640×192 resolution images on KITTI, then tested on Make3D. **Table 2** presents zero-shot transfer results. Under zero-shot settings, HyperDetailNet demonstrated strong generalization capability, outperforming most baseline methods. This proves our model architecture can effectively adapt to depth estimation tasks in different real-world scenarios.

4.4 Ablation Study

To verify the effectiveness of each component in HyperDetailNet, we conduct a series of ablation study. The results are shown in **Table 3**. We evaluate the impact of the dual-branch detail-global feature extraction DepthNet and the multi-attention fusion module on model performance.

The Benefit of Dual-Branch Detail-Global Feature Extraction DepthNet. As shown in **Table 3**, the complete HyperDetailNet achieves 0.118 in AbsRel, 0.839 in SqRel, and 4.748 in RMSE. When removing the dual-branch DepthNet and using only a single branch, performance drops significantly. Specifically, SqRel and RMSE errors increase by 10.1% and 3.6%, respectively. This result highlights the importance of the dual-branch detail-global feature extraction structure in preserving depth details.

The Benefit of Multi-Attention Fusion Module. When replacing the multi-attention fusion module with simple feature concatenation, performance declines further. The depth accuracy metric drops from 0.871 to 0.841, a 3.4% decrease. This demonstrates that different attention mechanisms can complement each other to improve depth estimation accuracy.

5 Conclusions

In this paper, we propose HyperDetailNet, a self-supervised monocular depth estimation network with a dual-branch DepthNet and multi-attention fusion. The dual-branch extraction of detail and global features helps resolve the issue of detail blurring in traditional methods. The proposed multi-attention fusion module integrates spatial attention, channel attention, and windowed self-attention to enhance and better fuse detail and global features. Experimental results show that HyperDetailNet achieves strong performance on the KITTI dataset. The generalization ability of the model is also validated on the Make3D dataset. Our work provides a compelling insight into the problems of blurred edges and missing details of depth maps in monocular depth estimation.

References

1. Chen, Y., Zhao, D., Lv, L., Zhang, Q.: Multi-task learning for dangerous object detection in autonomous driving. In: Information Sciences, 432, pp. 559-571 (2018)
2. Zhu, Z., Su, A., Liu, H., Shang, Y., Yu, Q.: Vision navigation for aircrafts based on 3D reconstruction from real-time image sequences. In: Science China Technological Sciences, 58, pp. 1196-1208 (2015)
3. Valentin, J., Kowdle, A., Barron, J. T., Wadhwa, N., Dzitsiuk, M., Schoenberg, M., et al.: Depth from motion for smartphone AR. In: ACM Transactions on Graphics (ToG), 37(6), pp. 1-19 (2018)
4. Saxena, A., Chung, S., Ng, A.: Learning depth from single monocular images. In: Advances in Neural Information Processing Systems, 18 (2005)
5. Rogers, B., Graham, M.: Motion parallax as an independent cue for depth perception. In: Perception, 8(2), pp. 125-134 (1979)
6. Ens, J., Lawrence, P.: An investigation of methods for determining depth from focus. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(2), pp. 97-108 (1993)
7. Garg, R., Bg, V. K., Carneiro, G., Reid, I.: 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, pp. 740-756. Springer International Publishing (2016)
8. Godard, C., Mac Aodha, O., Brostow, G. J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 270-279 (2017)
9. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems, 27 (2014)
10. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 239-248. IEEE (2016)
11. Bhat, S. F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4009-4018 (2021)
12. Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: Neural window fully-connected crfs for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3916-3925 (2022)

13. Zhou, T., Brown, M., Snavely, N., Lowe, D. G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1851-1858 (2017)
14. Godard, C., Mac Aodha, O., Firman, M., Brostow, G. J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3828-3838 (2019)
15. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5667-5675 (2018)
16. Feng, Z., Yang, L., Jing, L., Wang, H., Tian, Y., Li, B.: Disentangling object motion and occlusion for unsupervised multi-frame monocular depth. In: European Conference on Computer Vision, pp. 228-244. Cham: Springer Nature Switzerland (2022)
17. Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: Self-supervised multi-frame monocular depth. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1164-1174 (2021)
18. Zhou, Z., Fan, X., Shi, P., Xin, Y.: R-MSFM: Recurrent multi-scale feature modulation for monocular depth estimating. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12777-12786 (2021)
19. Yan, J., Zhao, H., Bu, P., Jin, Y.: Channel-wise attention-based network for self-supervised monocular depth estimation. In: 2021 International Conference on 3D Vision (3DV), pp. 464-473. IEEE (2021)
20. Song, L., Shi, D., Xia, J., Ouyang, Q., Qiao, Z., Jin, S., Yang, S.: Spatial-aware dynamic lightweight self-supervised monocular depth estimation. In: IEEE Robotics and Automation Letters, 9(1), pp. 883-890 (2023)
21. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. In: The International Journal of Robotics Research, 32(11), pp. 1231-1237 (2013)
22. Saxena, A., Sun, M., Ng, A. Y.: Make3d: Learning 3d scene structure from a single still image. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(5), pp. 824-840 (2008)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778 (2016)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, 30 (2017)
25. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012-10022 (2021)
26. Yang, Z., Wang, P., Wang, Y., Xu, W., Nevatia, R.: Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 691-709 (2018)
27. Ramamonjisoa, M., Lepetit, V.: Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 2109-2118 (2019)
28. Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., Ricci, E.: Structured attention guided convolutional neural fields for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3917-3925 (2018)
29. Johnston, A., Carneiro, G.: 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 (2020)

30. Zhang, F., Prisacariu, V., Yang, R., Torr, P. H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 185-194 (2019)
31. Klingner, M., Termöhlen, J. A., Mikolajczyk, J., Fingscheidt, T.: 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, pp. 582-600. Springer International Publishing (2020)
32. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234-241. Springer International Publishing (2015)
33. Woo, S., Park, J., Lee, J. Y., Kweon, I. S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19 (2018)
34. Li, Z., Bhat, S. F., Wonka, P.: Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10016-10025 (2024)
35. Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P.: Image quality assessment: from error visibility to structural similarity. In: IEEE Transactions on Image Processing, 13(4), pp. 600-612 (2004)
36. Zhang, N., Nex, F., Vosselman, G., Kerle, N.: 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 (2023)
37. Fan, D., Liu, S.: MonoRetNet: A Self-supervised Model for Monocular Depth Estimation with Bidirectional Half-Duplex Retention. In: International Conference on Intelligent Computing, pp. 361-372. Singapore: Springer Nature Singapore (2024)
38. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2650-2658 (2015)
39. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255. IEEE (2009)
40. Lyu, X., Liu, L., Wang, M., Kong, X., Liu, L., Liu, Y., et al.: Hr-depth: High resolution self-supervised monocular depth estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 3, pp. 2294-2301 (2021)
41. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, pp. 8001-8008 (2019)
42. Li, H., Gordon, A., Zhao, H., Casser, V., Angelova, A.: Unsupervised monocular depth learning in dynamic scenes. In: Conference on Robot Learning, pp. 1908-1917. PMLR (2021)
43. Zhang, C., Liu, J., Han, C.: Unsupervised learning of depth estimation based on attention model from monocular images. In: 2020 International Conference on Virtual Reality and Visualization (ICVRV), pp. 191-195. IEEE (2020)
44. Heydrich, T., Yang, Y., Du, S.: A lightweight self-supervised training framework for monocular depth estimation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2265-2269. IEEE (2022)

45. Cui, X. Z., Feng, Q., Wang, S. Z., Zhang, J. H.: Monocular depth estimation with self-supervised learning for vineyard unmanned agricultural vehicle. In: Sensors, 22(3), pp. 721 (2022)
46. Sun, Y., Hariharan, B.: Dynamo-depth: Fixing unsupervised depth estimation for dynamical scenes. In: Advances in Neural Information Processing Systems, 36, pp. 54987-55005 (2023)
47. Li, R., Yu, H., Du, K., Xiao, Z., Yan, B., Yuan, Z.: Adaptive Semantic Fusion Framework for Unsupervised Monocular Depth Estimation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE (2023)
48. Zhou, T., Brown, M., Snavely, N., Lowe, D. G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1851-1858 (2017)
49. Wang, C., Buenaposada, J. M., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2022-2030 (2018)
50. Xu, X., Chen, Z., Yin, F.: Multi-scale spatial attention-guided monocular depth estimation with semantic enhancement. In: IEEE Transactions on Image Processing, 30, pp. 8811-8822 (2021)