IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 June 2024, accepted 30 August 2024, date of publication 3 September 2024, date of current version 11 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3453942

RESEARCH ARTICLE

Monocular Depth Estimation Based on Residual Pooling and Global-Local Feature Fusion

LINKE LI[1], ZHENGYOU LIANG[1,2], XINYU LIANG[1], AND SHUN LI[1]
[1]School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China
[2]Guangxi Key Laboratory of Multimedia Communications and Network Technology, Nanning 530004, China

Corresponding author: Zhengyou Liang (zhyliang@gxu.edu.cn)

This work was supported by the National Nature Science Foundation of China under Grant 62171145.

ABSTRACT To improve the prediction accuracy of monocular depth estimation networks and address issues such as edge blurring and excessive artifacts in the generated depth maps, this paper proposes a deep network architecture based on a global-local feature fusion module and a residual pooling module. The encoder utilizes a Hierarchical Transformer, while the decoder incorporates a U-Net structure model that combines multi-dimensional attention features aggregation and residual pooling. The residual pooling module facilitates better extraction of background contextual information from the feature maps to obtain more precise scene depth information. The global-local feature fusion module enables the network to learn features that encompass both global and local information. Experimental evaluations conducted on the NYU Depth V2 and KITTI datasets demonstrate that the proposed method achieves a $\delta 1$ of 0.916 on the NYU Depth V2 dataset, along with enhanced generalization ability and robustness. Furthermore, the effectiveness of each module is validated through ablation studies on the NYU Depth V2 dataset.

INDEX TERMS Computer vision, convolutional neural networks, image processing, monocular depth estimation, residual pooling.

I. INTRODUCTION

Monocular depth estimation is the technology of predicting the distance (depth) from the scene to the center of the camera imaging using a single RGB image and generating a depth image containing depth information. A technology has wide application prospects in 3D reconstruction [1], robotics [2], and autonomous driving [3]. However, monocular depth estimation faces the challenge of uncertainty in image scale, as the same input image can correspond to numerous different 3D scenes. To address this issue, researchers typically adopt two approaches. One approach is to use real depth maps as supervision for training and evaluation of depth, while the other approach involves learning the depth structure and geometric information of the scene from a single image to achieve depth estimation.

Following the pioneering work of Saxena et al. [4] in using segmented plane structures in the scene and Markov random fields to predict depth, the academic community began studying supervised learning-based monocular depth estimation. Subsequent research introduced methods such as conditional random fields (CRF) [5], [6], modified loss functions [2], and multi-scale feature fusion [3]. Unsupervised depth estimation does not require depth labels or other supervision information. Godard et al. [7] proposed Monodepth, which performs depth estimation by learning the ego-motion and depth information of monocular images, capable of handling cases with texture scarcity and high scene complexity. Subsequently, Godard et al. [8] introduced Monodepth2, which uses minimal reprojection loss to alleviate occlusion issues and employs an auto-masking loss to filter out moving objects with the same speed as the camera.

As research progresses, researchers have gradually recognized the importance of global and local information in generating high-quality depth images [9]. Therefore, there is a need for the ability to comprehend the global scene while fully utilizing local information.

To further improve the accuracy of depth estimation and generate depth maps with clear boundaries, this paper proposes a new network architecture. Firstly, Hierarchical

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

Transformer is built as an encoder to construct global relationships to efficiently capture multi-scale context features. Subsequently, a U-Net structured model is designed in the decoder, incorporating the Aggregated Multi-Dimensional Attention Feature Module (MAMF) and Residual Pooling Module (RPM). The MAMF adaptively adjusts the weights of various channels and spatial positions in the feature map through attention mechanisms to extract pertinent information. Following this, the Global-Local Feature Fusion Module (GLFFM) is utilized to construct multi-scale feature concatenation paths combining the network's global and local features. However, using only MAMF and GLFFM module-designed decoder layers with shallow depth can lead to inadequate decoding of depth information in certain scenes. To address this issue, the RPM is introduced, which retains information from the original input during the learning process while reducing the number of parameters. Skip and cross-stage connections are also employed to better interconnect these modules. To validate the effectiveness of the proposed network architecture, extensive experiments are conducted on the NYU Depth V2 [10] and KITTI [11] datasets. Experimental results demonstrate that the proposed network model improves depth accuracy and boundary prediction accuracy with only 61M parameters. The effectiveness of the model and each module is qualitatively and quantitatively verified, showing good generalization and robustness in cross-dataset testing.

The main contributions of this paper can be summarized as follows:

1. A Global-Local Feature Fusion Module was designed to enhance edge information and depth prediction accuracy by effectively utilizing the global and local feature information of the depth map.

2. A Residual Pooling Module was designed, utilizing max-pooling to capture background information of the depth image and employing residual methods to ensure that the output information is not solely dependent on the network's learning capacity.

3. The effectiveness of the proposed method was experimentally validated on the NYU Depth V2 [10] and KITTI [11] datasets, exhibiting greater robustness compared to previous networks. Additionally, excellent generalization capabilities were demonstrated on the indoor dataset SUN RGB-D [12].

II. RELATED WORK

The choice of network architecture in monocular depth estimation directly influences the model's ability to accurately infer depth, particularly in how it processes and integrates spatial and contextual information. Most of the research in monocular depth estimation is based on encoder-decoder network architectures. For the design of the encoder, features are primarily extracted through convolutional neural networks (CNN), Transformers, or a hybrid of both methods.

Eigen et al. [1] were the first to apply CNN to monocular depth estimation research, proposing a network structure that combines coarse-scale and fine-scale depth predictions to generate dense pixel-wise depth estimates. This research laid the foundation for subsequent studies in monocular depth estimation and advanced the field. Early studies primarily utilized convolutional neural networks based on VGG [12]; however, the VGG network faces issues of parameter redundancy and high computational complexity. Yin and Shi [14] achieved state-of-the-art performance in monocular depth estimation research by replacing the VGG model's network architecture with ResNet [15]. ResNet tackles the problem of vanishing and exploding gradients by introducing residual connections, addressing the issues present in the VGG model. Laina et al. [2] utilized ResNet as the encoder and proposed a fully convolutional residual network with BerHu loss [16], enabling the network to generate dense pixel-wise depth maps with improved resolution in simpler scenarios. Lee et al. [17] introduced a novel network architecture utilizing multiple stages of novel local planar guidance layers in the decoding stage for full-resolution depth estimation. Kumari et al. [18] incorporated residual connections and introduced perceptual loss in the encoder-decoder CNN network architecture, considering high-level features at different scales to aid the model in faster convergence.

Transformer can replace traditional convolutional neural networks for feature extraction. The self-attention mechanism with multi-layer perceptrons in Transformer allows for modeling global relationships, enabling more accurate capturing of long-range dependencies in images. Bhat et al. [19] proposed a structure combining a CNN-based encoder and Transformer blocks to divide the depth range into adaptive estimated center values for each image, converting the depth estimation task into a classification task and demonstrating state-of-the-art performance. Shao et al. [20] introduced uncertainty-aware cross distillation between Transformer and CNN as an encoder, leveraging Transformer branches for establishing long-range correlations and CNN branches for focusing on local information. Gordon et al. [21] presented an unsupervised monocular depth estimation method based on the Transformer network architecture, effectively integrating and fusing features of different scales and resolutions.

Following the success of the Vision Transformer [22] in image classification tasks, researchers have widely applied it in the field of monocular depth estimation in recent years. Ranftl et al. [23] utilized a network architecture based on Vision Transformer-CNN as an encoder, providing more fine-grained and globally consistent predictions, but the model training requires large datasets and takes longer training times. Zheng et al. [24] demonstrated the advantage of Vision Transformer in dense prediction tasks. Subsequently, Xie et al. [25] proposed the SegFormer model, a Transformer-based segmentation framework with a simple, lightweight MLP decoder, better suited for application in monocular depth estimation. Wu and Wang [26] achieved promising results in monocular depth estimation using the SegFormer model; however, the large model parameters may lead to poor generalization capabilities. Lie et al. [27] also employed

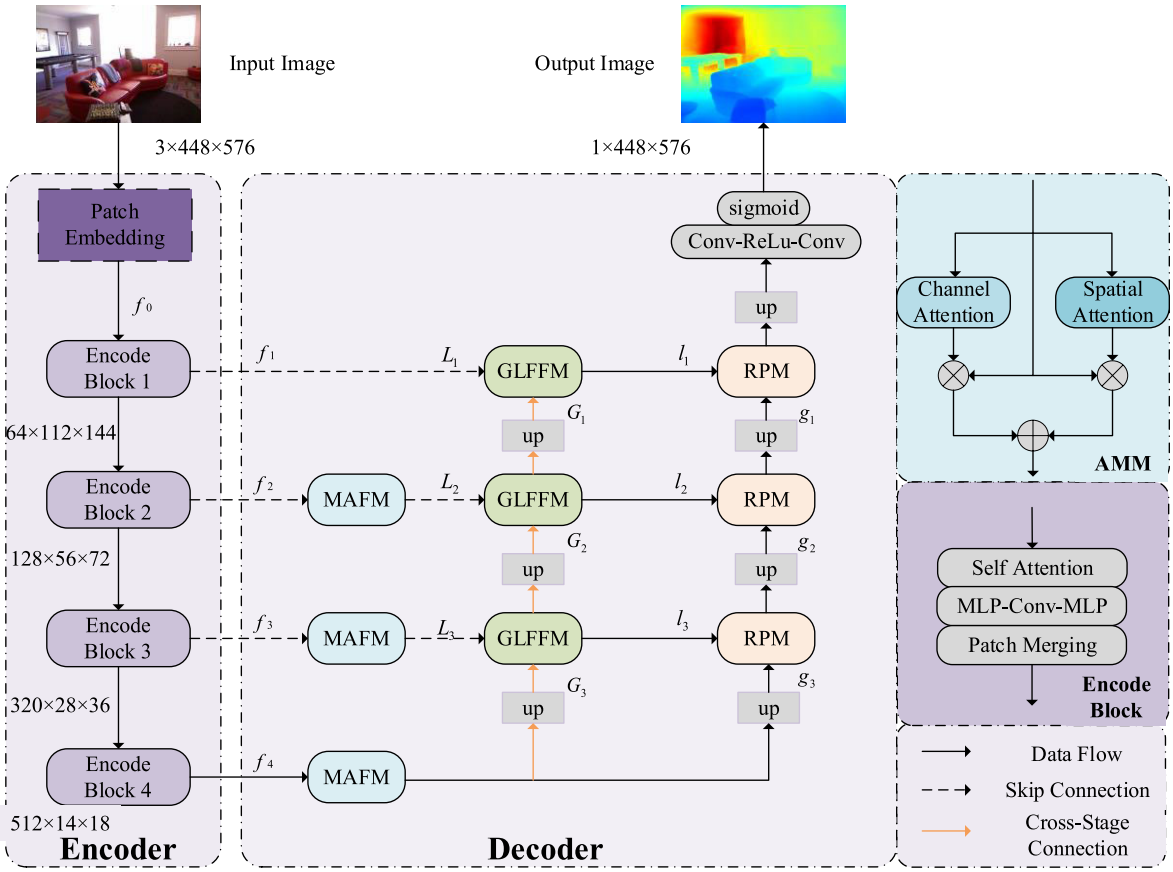


FIGURE 1. Overview of the proposed network architecture. The framework of our network is the encoder-decoder design. The Vision Transformer is considered as the encoder, with which sufficient multi-level global features are extracted. In the decoder, these modules guide the network to better predict full resolution depth maps with clear structure details.

the SegFormer model, but the shallow decoder layers could result in inadequate depth information decoding for certain scenarios.

In summary, the CNN encoder-decoder network architecture exhibits inaccuracies in depth prediction, low precision in scene boundary predictions, and large model parameters. Conversely, utilizing a combination of CNN and Transformer- based encoders increases computational complexity, affecting training speed and leading to significant memory overhead. Therefore, we propose using the more efficient Hierarchical Transformer as an encoder and appropriately widening the decoder to focus on rich global and local features in feature maps, aiming to achieve more precise depth estimation performance.

III. METHODOLOGY

The purpose of the depth estimation model is to accurately predict the depth map $O \in R^{3\times w\times h}$ of a given RGB image $I \in R^{1\times w\times h}$. Our depth estimation network consists of an encoder and a decoder, and the specific depth estimation model is shown in Figure 1.

Encoder: We use the SegFormer [25] model as the encoder due to its efficiency in capturing multi-scale contextual features through a series of Transformer-based blocks. The encoder of the SegFormer model is composed of a series of Transformer-based blocks, with each Transformer Block connected through residual connections and layer normalization operations. Additionally, it reduces the computational cost of self-attention mechanisms through Shifted Window and utilizes the Patch Embedding [22] method to divide the input image into multiple small blocks for encoding. Initially, the input image undergoes Patch Embedding processing to obtain feature maps f_0, which are then sequentially input into the Transformer Block to obtain feature maps f_1, f_2, f_3, and f_4. Their corresponding channel numbers and image sizes are f_1: 64 × 112×144, f_2: 128 × 56×72, f_3: 320 × 28×36, f_4: 512 × 14×18. Simultaneously, during the encoding process, it is capable of generating multi-scale features for use in the decoding stage.

Decoder: To achieve accurate depth map estimation results, we construct a novel decoder structure based on the U-Net architecture to restore features to a size of $1 \times H \times W$. Additionally, to fully utilize the high-level semantic features extracted by the encoder, we increase the width of the decoder. In this paper, we first adaptively adjust the weights of various channels and spatial positions in the feature maps

through the MAMF module to reduce the number of channels. Subsequently, by stacking and concatenating the GLFFM and RPM sub-modules, we gradually extract features at different levels and perform scale feature fusion to better capture multi-scale and multi-level features of the target. The designed decoder can focus on both global and local feature information separately, thereby improving the accuracy and robustness of depth estimation. Local feature information is achieved through inter-stage connections, establishing connections between feature maps at different stages to facilitate information transfer and fusion. Additionally, the proposed sub-modules need to construct decoding paths through skip connections to more accurately restore the target depth map. Section III-A, III-B and III-C provide a detailed introduction to these modules.

A. MULTI-DIMENSIONAL ATTENTION FEATURE MODULE (MAFM)

The Multi-Dimensional Attention Feature Aggregation Module amplifies the respective advantages of channel attention and spatial attention by combining them. It mainly leverages the channel information and spatial positional information of feature maps and is illustrated in the model architecture diagram shown in Figure 2.

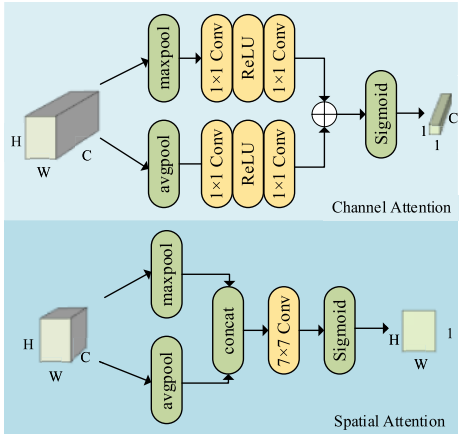


FIGURE 2. Multi-dimensional attention feature module.

Channel attention is used to enhance the channel correlation of feature maps. It acquires background information of the channels through average pooling and obtains attention weights for each channel using a sigmoid activation function. Finally, these attention weights are multiplied by the input feature map to enhance the representation capacity of important channels. Spatial attention, on the other hand, is utilized to enhance the spatial correlation of feature maps. The attention weights obtained through a sigmoid activation function are multiplied by the input feature map to enhance the representation capacity of important spatial positions. The entire process can be represented by the following formula:

$$x_{channel} = x_n \times Sigmoid(MLP(\max pool(F))). \quad (1)$$

$$x_{spatial} = x_n \times Sigmoid(W^{7\times7}([\max pool(F); avgpool(F)])). \quad (2)$$

$$x_{out} = x_{channel} + x_{spatial}. \quad (3)$$

where F is the input feature map, x_{in} is the feature value at the corresponding position on the input feature map F, MLP is 1×1 Conv-ReLU-1×1Conv, and $M^{7\times7}$ represents the weight matrix of the 7×7Conv.

B. GLOBAL-LOCAL FEATURE FUSION MODULE (GLFFM)

The Global-Local Feature Fusion Module leverages information related to image depth in both the global pathway and local pathway, rather than focusing excessively on texture information of objects in the image. The model structure diagram is shown in Figure 3.

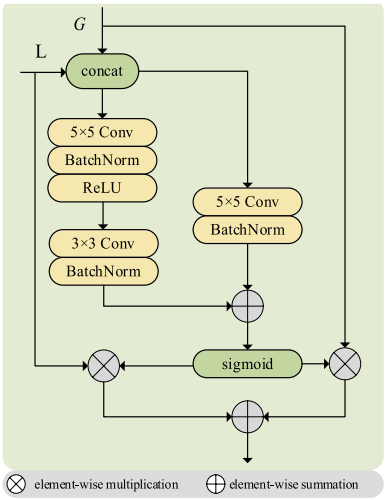


FIGURE 3. Global-local feature fusion module.

The global features and local features are concatenated using the concat operation. Subsequently, the feature map from the 5×5 Conv-Batch Normalization-ReLU layer and the feature map from the 3×3-Batch Normalization layer are added to the result from the 5×5-Batch Normalization layer to obtain a dual-channel feature map. Finally, the global features are added to the local features multiplied by the results of each channel, emphasizing the important locations in the feature map. The 5×5 convolutional kernel can capture broader contextual information, while the 3×3 convolution focuses on details and local information. This module helps refine global and local feature information, effectively utilizing the depth information of the image and enhancing the efficiency of the decoder.

C. RESIDUAL POOLING MODULE (RPM)

The Residual Pooling Module can preserve the most significant features and output fused feature maps with rich spatial information, enabling the network to extract and learn features more effectively in subsequent layers. The model structure diagram is shown in Figure 4.

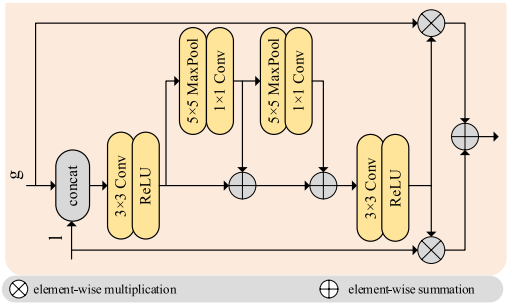


FIGURE 4. Residual pooling module.

This module consists of two pooling blocks, each containing a max pooling layer and a convolutional layer. The convolutional layer acts as a weighted fusion operation, learning weights to adapt to the pooling blocks during the training process. Alternating between max pooling and convolution operations can enhance the model's representational capacity, reduce the number of parameters, and improve computational efficiency. The output feature maps of all pooling blocks are fused with the input feature map through residual connections by summation. This allows the network to retain information from the original input during learning, rather than solely relying on the network's learning capacity. The specific computational process is as follows:

$$x_i = C(P(x_{i-1})), i = 1, 2. \tag{4}$$

$$x_{out} = \sum_{x=0}^{2} x_i. \tag{5}$$

$$X_{out} = x_{local} \times R(C_3(x_{out})) + x_{global} \times R(C_3(x_{out})) \tag{6}$$

where C represents 1×1Conv, P represents 5×5 maxpool, R represents ReLU, C_3 represents 3×3Conv, and x_0 represents the feature after being processed by 3×3Conv-ReLU.

D. DATA AUGMENTATION

Data augmentation is an effective technique for alleviating overfitting by increasing the effective number of training samples. In the field of monocular depth estimation, methods such as CutDepth [28] and CutFlip [20] are commonly used for data augmentation. The CutDepth method replaces parts of the input image with ground truth depth maps, providing multiple possibilities for the input image and allowing the network to focus on high-frequency components. On the other hand, the CutFlip method vertically splits the input image into two parts and flips them along the vertical direction, in order to weaken the correlation between depth and the vertical position of the image.

However, monocular depth estimation models heavily rely on vertical image positions [29] to infer depth, often overlooking other cues such as apparent size, which can lead to reduced model generalization. In order to prevent overfitting and enable the network to focus on high-frequency regions, we employed the CutDepth method along with a combination of random data augmentation techniques, including a 50% probability of horizontal flipping, random brightness, random gamma transformation, and adjustments to hue and saturation values. It should be noted that color augmentation is only applied to the input images and not to the synthetic images.

E. TRAINING LOSS

To facilitate the computation of the error between the predicted depth map Y^* and the ground truth depth map Y, we adopt the scale-invariant logarithmic scale loss 0 to train the model. This method has advantages such as fast convergence, high prediction quality, and low loss. Specifically, the training loss equation can be expressed as:

$$loss = \frac{1}{n}\sum_i d_i^2 - \frac{1}{2n^2}\sum_i d_i^2 \tag{7}$$

where $d_i = logy_i - logy_i^*$, y_i, and y_i^* represent the i-th pixel in Y and Y^*.

IV. EXPERIMENT

To substantiate the model's performance, we performed comprehensive experiments and evaluations on both the challenging indoor NYU Depth V2 and diverse outdoor KITTI datasets. We compared our model with existing methods through quantitative and qualitative assessments, and performed ablation experiments to verify the effectiveness of each module's contribution.

A. DATASET

NYU Depth V2 [10] is an indoor dataset collected by Microsoft Kinect's RGB and Depth cameras, with a resolution of 640 × 480 for indoor images and a depth range of 0 to 10 meters. We follow the official training/testing segmentation method to evaluate the performance of the model, using approximately 24K images randomly cropped to 576 × 448 for network training, and evaluated on 654 images.

KITTI [11] is the most commonly used dataset for capturing outdoor scenes from moving vehicles, with images having a resolution of 352 × 1216 and depth values ranging from 0 to 80 meters. We follow the data segmentation method of Eigen et al. [1], using approximately 23K training set images randomly cropped to 704 × 352 for network training, and evaluated on 697 test set images.

B. EVALUATION

We adopt five common evaluation metrics [1] used in monocular depth estimation to compare our method with state-of-the-art models. Among them, n represents the total number of obtainable pixels in the ground truth depth map, d_p represents the predicted depth given at pixel p, and d_p^* represents the true depth at pixel p. $|.|$ returns the number of elements in the input set. The evaluation metrics are defined as follows:

$$REL = \frac{1}{n}\sum_{p=1}^{n} \frac{\left|d_p - d_p^*\right|}{d_p^*} \tag{8}$$

TABLE 1. Accuracy evaluation of different methods tested on NYU Depth V2 dataset.

Method	Params(M)	Higher is better(δ<) ↑			Lower is better ↓		
		1.25^1	1.25^2	1.25^3	REL	RMSE	log10
Xue et al. [31]	220	0.846	0.969	0.992	0.123	0.550	0.053
Lee et al. [17]	47	0.885	0.978	0.994	0.110	0.392	0.047
Patil et al. [32]	-	0.898	0.981	0.996	0.104	0.356	0.043
Agarwal et al. [33]	-	0.900	0.983	0.996	0.106	0.365	0.045
Bhat et al. [19]	78	0.903	0.984	0.997	0.103	0.364	0.044
Ranftl et al. [23]	123	0.904	**0.988**	0.997	0.110	0.357	0.045
Kim et al.* [9]	62	0.912	0.986	0.996	0.099	0.346	0.042
Lei et al. [27]	60	0.912	0.986	**0.998**	**0.098**	0.344	0.043
Wu et al. [26]	67	0.914	0.987	0.996	0.099	0.347	0.042
Ours	61	**0.916**	**0.988**	0.997	**0.098**	**0.334**	**0.041**

Note: bold indicates the best result, underline indicates the second best result, * indicates the result obtained in our experimental environment

$$RMSE = \sqrt{\frac{1}{n}\sum_{p=1}^{n}\frac{\left|d_p - d_p^*\right|^2}{d_p^*}} \tag{9}$$

$$RMSE\log = \sqrt{\frac{1}{n}\sum_{p=1}^{n}\left|\lg d_p - \lg d_p^*\right|^2} \tag{10}$$

$$\delta_i = \frac{1}{n}\left|\left\{d \in n \middle| \max\left(\frac{d_p^*}{d_p}, \frac{d_p}{d_p^*}\right) < 1.25^t\right\}\right| \times 100\% \tag{11}$$

C. IMPLEMENTATION DETAIL

This experiment uses the PyTorch framework to conduct experiments on an NVIDIA TITAN Xp GPU with 12GB of memory. The training epoch is set to 25, and the batch size is set to 4, with the same settings applied to both the NYU Depth V2 and KITTI datasets.

To train the network, we use a single-cycle learning rate strategy with the Adam optimizer [30]. The learning rate is dynamically adjusted using an LR scheduler, with a decay exponent of 0.9 to slow down its growth. We set the learning rate range from 3e-5 to 1e-4, and calculate the current learning rate using linear interpolation based on the model's training progress (the ratio of global steps to total steps) and exponential decay. The learning rate is adjusted between the minimum and maximum values, with a tendency to increase first and then decrease as training progresses. This learning rate adjustment strategy can improve the training effect and convergence speed of the model.

D. COMPARISON WITH STATE OF THE ARTS

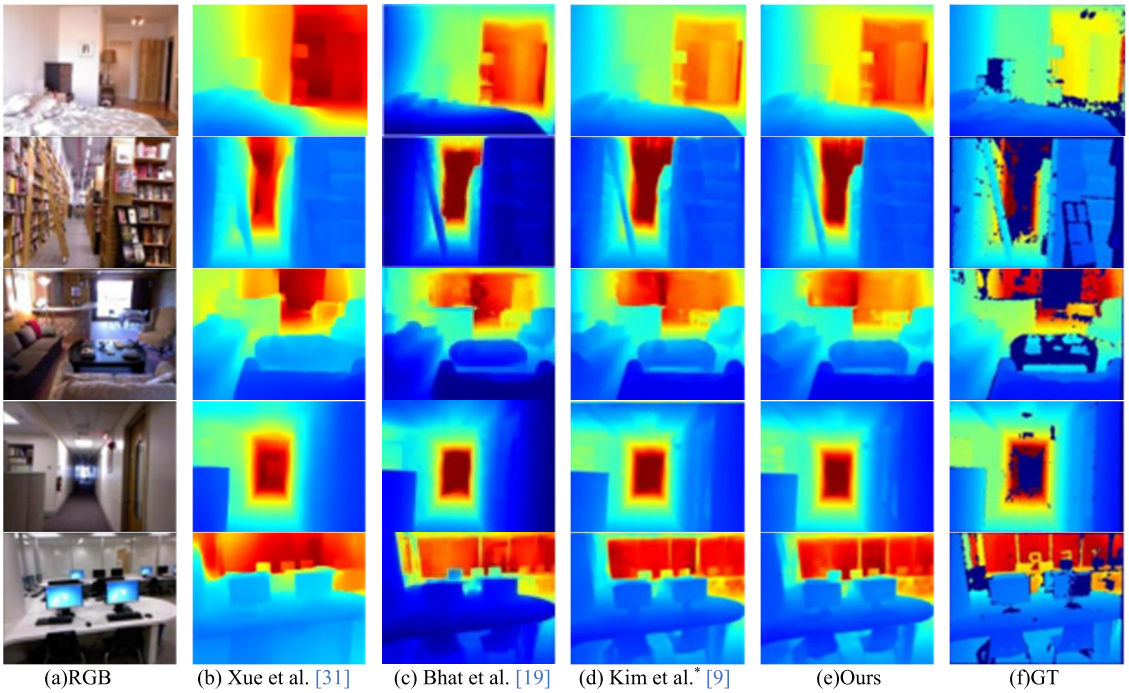
NYU Depth V2: Table 1 presents the comparison of our proposed method with some state-of-the-art methods on the NYU Depth V2 dataset. ''Params'' represents the model's parameter count, the third column shows the pixel-level accuracy of the predicted depth maps at thresholds δ_i of 1.25^1, 1.25^2, and 1.25^3, where higher values indicate better performance. The fourth column presents three error metrics, where lower values indicate better performance. From the table, it can be observed that our proposed model has only 61M parameters. Additionally, five out of the six metrics demonstrate good performance. Xue et al. [31] used the Sobel operator for edge detection to induce boundary information generation, leading the model to focus on boundary information while neglecting overall depth information. Lee et al. [17] introduced multi-scale local plane guidance, requiring a large amount of training data. Patil et al. [32] proposed a method based on segmented plane priors, which may suffer from generalization issues in complex scenes. Patil et al. [32] introduced a monocular depth estimation method based on a multi-scale vision transformer, which might increase the computational complexity of training and inference. Bhat et al. [19] partition depth values during depth estimation, which can affect the accuracy of the depth estimation. Ranftl et al. [23] require a large number of parameters, increasing the model's computational cost. Kim et al. [9] utilized a lightweight decoder, which may result in a shallow decoder unable to meet the demands of different scenes. Lei et al. [27] may exhibit inaccurate boundary information predictions in scenes such as mirrors and doors. Wu and Wang [26] lacked experiments on the robustness and generalization ability of the model. Among all the compared methods, our proposed model demonstrates advanced performance in most evaluation metrics, which we attribute to the proposed architecture, enabling better extraction of depth information from images.

Figure 5 presents the visual results of the NYU Depth V2 dataset in five different indoor scenes (bedroom, bookstore, living room, office, and computer lab). By comparing the predicted depth maps, it can be observed that our method produces clearer and more reasonable results than those reported in the literature. Our method effectively reduces depth map artifacts in distant locations such as lamps, tables, chairs, and door frames, resulting in sharper boundaries

TABLE 2. Accuracy evaluation of different methods tested on KITTI dataset.

Method	Params(M)	Higher is better(δ<) ↑			Lower is better ↓		
		1.25^1	1.25^2	1.25^3	REL	RMSE	RMSE log
Fu et al. [3]	110	0.932	0.984	0.994	0.072	2.727	0.120
Yin et al. [34]	114	0.938	0.984	0.998	0.072	3.258	0.117
Lee et al. [17]	113	0.956	0.993	0.998	0.059	2.756	0.088
Ranftl et al. [23]	123	0.959	0.995	**0.999**	0.062	2.573	0.092
Bhat et al. [19]	78	0.964	0.995	**0.999**	0.058	2.360	0.088
Kim et al.* [9]	62	**0.966**	0.995	**0.999**	0.057	2.354	**0.087**
Ours	**61**	**0.966**	**0.996**	**0.999**	**0.056**	**2.315**	**0.087**

Note: bold indicates the best result, underline indicates the second best result, * indicates the result obtained in our experimental environment



(a)RGB (b) Xue et al. [31] (c) Bhat et al. [19] (d) Kim et al.* [9] (e)Ours (f)GT

FIGURE 5. Visualise the results of different methods on different indoor scenes on NYU Depth V2 dataset. * indicates the result obtained in our experimental environment.

and better recovery of the depth information from the images.

KITTI: To demonstrate the effectiveness and superiority of the proposed method in this paper, we compared our network model with recent relevant methods on the KITTI dataset, and the results are shown in Table 2. It can be observed from the table that our proposed method achieves advanced performance in all six metrics on the KITTI dataset.

SUN RGB-D: To evaluate the network's generalization ability, we conducted additional testing on the indoor SUN RGB-D dataset. We trained the network using the NYU Depth V2 dataset and evaluated it on the SUN RGB-D test set. Figure 6 presents a qualitative comparison of different networks on the SUN RGB-D test set. The results indicate that our proposed network performs better in depth estimation for distant scenes, and the boundaries of the generated depth maps are also clearer. This demonstrates the superiority of our method and its better generalization ability.

E. ROBUSTNESS EXPERIMENTS

Robustness experiments are an effective method for evaluating the performance of algorithms in the presence of noise, changes in lighting, occlusions, motion blur, and other conditions. They can also assess the stability and reliability of algorithms in different environments. To verify the adaptability of the proposed network under different conditions and scenarios, we conduct robustness experiments. We train the network using the NYU Depth V2 dataset and then test it on the NYU Depth V2 test set with the addition of Gaussian noise, motion blur, changes in contrast, and snow, as shown in Table 3.

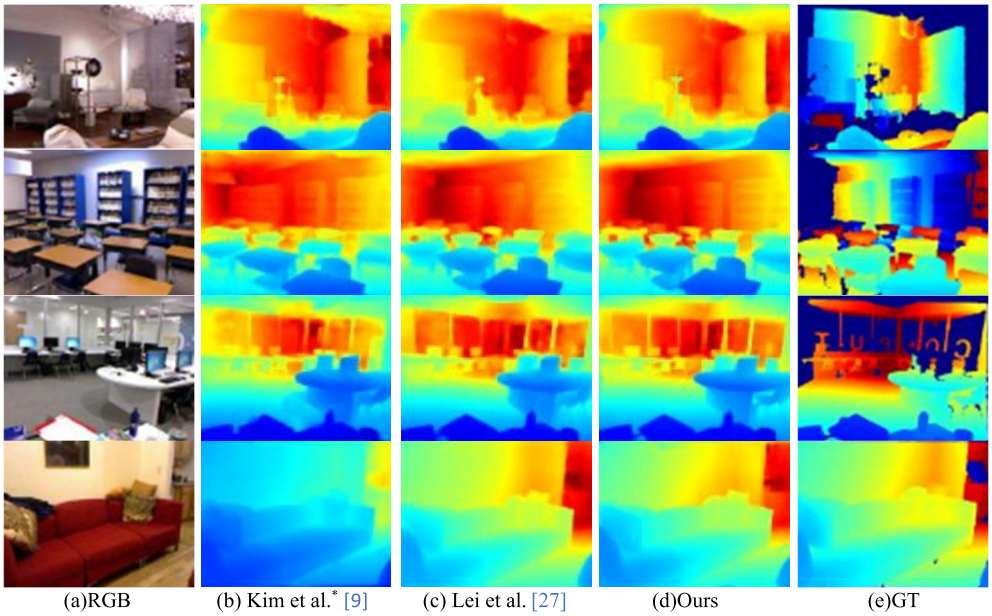


FIGURE 6. Visualise the results of different methods on different indoor scenes on SUN-RGBD dataset. * indicates the result obtained in our experimental environment.

TABLE 3. Robustness experiment results on corrupted images of NYU Depth V2 dataset.

		Higher is better(δ<) ↑			Lower is better ↓		
Corruption Type	**Method**	**1.25^1**	**1.25^2**	**1.25^3**	**REL**	**RMSE**	**RMSElog**
Gaussian Noise	Lee et al. [17]	0.223	0.384	0.543	0.435	1.589	0.743
	Bhat et al. [19]	0.347	0.553	0.708	0.343	1.299	0.544
	Kim et al. [9]	0.775	0.940	0.983	0.161	0.541	0.198
	Ours	**0.825**	**0.962**	**0.991**	**0.143**	**0.459**	**0.172**
Motion Blur	Lee et al. [17]	0.667	0.850	0.922	0.189	0.701	0.279
	Bhat et al. [19]	0.697	0.859	0.927	0.180	0.643	0.262
	Kim et al. [9]	0.807	0.946	0.981	0.139	0.494	0.183
	Ours	**0.893**	**0.983**	**0.996**	**0.108**	**0.372**	**0.135**
Contrast	Lee et al. [17]	0.697	0.864	0.932	0.181	0.689	0.263
	Bhat et al. [19]	0.654	0.836	0.917	0.198	0.752	0.283
	Kim et al. [9]	0.860	0.971	0.992	0.117	0.427	0.152
	Ours	**0.910**	**0.986**	**0.996**	**0.100**	**0.350**	**0.126**
Snow	Lee et al. [17]	0.410	0.649	0.803	0.298	1.114	0.458
	Bhat et al. [19]	0.410	0.656	0.817	0.292	1.094	0.440
	Kim et al. [9]	0.723	0.926	0.981	0.170	0.598	0.217
	Ours	**0.901**	**0.982**	**0.996**	**0.107**	**0.366**	**0.133**

Note: bold indicates the best result

Observing Table 3, it can be seen that our method still exhibits strong depth estimation capability when applied to images with added Gaussian noise, motion blur, changes in contrast, and snow, demonstrating a certain level of robustness.

F. ABLATION EXPERIMENTS

This paper improves the method based on the baseline approach in reference [9]. To validate the effectiveness of each module, we conduct ablation experiments by systematically removing components. In this section,

TABLE 4. Ablation study on model architectures. All the models are trained and tested on NYU Depth V2 dataset.

Method	Higher is better(δ<) ↑			Lower is better ↓		
	1.25^1	1.25^2	1.25^3	REL	RMSE	log10
Baseline	0.911	0.986	0.996	0.099	0.346	0.042
Baseline-GLFFM	0.911	0.987	0.997	0.100	0.348	0.042
Baseline-RPM	0.912	0.986	0.996	0.099	0.342	0.042
Baseline-GLFFM-RPM	0.914	0.987	0.997	0.098	0.336	0.042
Baseline-GLFFM-RPM-MAFM	**0.916**	**0.988**	**0.997**	**0.098**	**0.334**	**0.041**

Note: bold indicates the best result

we experimentally verify the importance of the design choices for each module in the model by conducting experiments on the NYU Depth V2 dataset, and the results are presented in Table 4.

Observing Table 4, it can be seen that the addition of the GNFFM and RPM modules to the decoder part of the baseline has resulted in improved accuracy, with a decreasing trend in error metrics. This indicates that the introduction of these two modules in this paper has significantly contributed to the improvement of depth estimation accuracy. Subsequently, with the addition of the MAFM module, it can be observed that, except for δ_3 and REL, all other metrics have shown a slight improvement.

V. CONCLUSION

This paper demonstrates outstanding performance on the KITTI and NYU Depth V2 dataset by introducing MAFM, GNFFM, and RPM modules, representing a significant improvement compared to the original network. Specifically, the GNFFM and RPM modules can more accurately predict distant objects, possess superior global feature processing capabilities, and better preserve object details and contour information in depth maps. However, limitations exist in accurately predicting reflective objects such as mirrors and glass, particularly in terms of boundary prediction accuracy. Additionally, the proposed method sets the channel number for each feature layer during design, which may not necessarily be the optimal solution for different scenarios. Therefore, future research will focus on addressing these issues to achieve superior results.

REFERENCES

[1] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014, *arXiv:1406.2283*.

[2] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Stanford, CA, USA, Oct. 2016, pp. 239–248, doi: 10.1109/3DV.2016.32.

[3] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 2002–2011, doi: 10.1109/CVPR.2018.00214.

[4] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, May 2009, doi: 10.1109/TPAMI.2008.132.

[5] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1119–1127, doi: 10.1109/CVPR.2015.7298715.

[6] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 161–169, doi: 10.1109/CVPR.2017.25.

[7] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6602–6611, doi: 10.1109/CVPR.2017.699.

[8] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3827–3837, doi: 10.1109/ICCV.2019.00393.

[9] D. Kim, W. Ka, P. Ahn, D. Joo, S. Chun, and J. Kim, "Global–local path networks for monocular depth estimation with vertical CutDepth," 2022, *arXiv:2201.07436*.

[10] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 7576, Berlin, Germany, 2012, pp. 746–760, doi: 10.1007/978-3-642-33715-4_54.

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013, doi: 10.1177/0278364913491297.

[12] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 567–576, doi: 10.1109/CVPR.2015.7298655.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[14] Z. Yin and J. Shi, "GeoNet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 1983–1992, doi: 10.1109/CVPR.2018.00212.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[16] L. Zwald and S. Lambert-Lacroix, "The BerHu penalty and the grouped effect," 2012, *arXiv:1207.6868*.

[17] J. Han Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," 2019, *arXiv:1907.10326*.

[18] S. Kumari, R. R. Jha, A. Bhavsar, and A. Nigam, "AUTODEPTH: Single image depth map estimation via residual CNN encoder–decoder and stacked hourglass," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Taipei, Taiwan, Sep. 2019, pp. 340–344, doi: 10.1109/ICIP.2019.8803006.

[19] S. Farooq Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 4008–4017, doi: 10.1109/CVPR46437.2021.00400.

[20] S. Shao, Z. Pei, W. Chen, R. Li, Z. Liu, and Z. Li, "URCDC-depth: Uncertainty rectified cross-distillation with CutFlip for monocular depth estimation," 2023, *arXiv:2302.08149*.
[21] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8976–8985, doi: 10.1109/ICCV.2019.00907.
[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[23] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 12159–12168, doi: 10.1109/ICCV48922.2021.01196.
[24] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 6877–6886, doi: 10.1109/CVPR46437.2021.00681.
[25] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," 2021, *arXiv:2105.15203*.
[26] B. Wu and Y. Wang, "Rich global feature guided network for monocular depth estimation," *SSRN Electron. J.*, doi: 10.2139/ssrn.4057946.
[27] C. Lei, L. Zhengyou, and S. Yu, "A monocular image depth estimation method based on weighted fusion and point-wise convolution," *IET Comput. Vis.*, vol. 17, no. 8, pp. 1005–1016, 2023, doi: 10.1049/cvi2.12212.
[28] Y. Ishii and T. Yamashita, "CutDepth: Edge-aware data augmentation in depth estimation," 2021, *arXiv:2107.07684*.
[29] T. Van Dijk and G. De Croon, "How do neural networks see depth in single images?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 2183–2191, doi: 10.1109/ICCV.2019.00227.
[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[31] F. Xue, J. Cao, Y. Zhou, F. Sheng, Y. Wang, and A. Ming, "Boundary-induced and scene-aggregated network for monocular depth prediction," *Pattern Recognit.*, vol. 115, Jul. 2021, Art. no. 107901, doi: 10.1016/j.patcog.2021.107901.
[32] V. Patil, C. Sakaridis, A. Liniger, and L. Van Gool, "P3Depth: Monocular depth estimation with a piecewise planarity prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 1600–1611, doi: 10.1109/CVPR52688.2022.00166.
[33] A. Agarwal and C. Arora, "Depthformer: Multiscale vision transformer for monocular depth estimation with global local information fusion," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Bordeaux, France, Oct. 2022, pp. 3873–3877, doi: 10.1109/ICIP46576.2022.9897187.
[34] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 5683–5692, doi: 10.1109/ICCV.2019.00578.

LINKE LI is currently pursuing the master's degree with Guangxi University, Nanning, China. Her research interest includes monocular depth estimation in computer vision.

ZHENGYOU LIANG was born in 1968. He received the Ph.D. degree. He is currently a Professor. His main research interests include computer vision, artificial intelligence, and parallel distributed computing. He is a member of CCF.

XINYU LIANG is currently pursuing the master's degree with Guangxi University, Nanning, China. Her research interest includes object detection in computer vision.

SHUN LI is currently pursuing the master's degree with Guangxi University, Nanning, China. His research interest includes monocular depth estimation in computer vision.