frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 12 January 2023
DOI 10.3389/fphy.2022.1115764

OPEN ACCESS

EDITED BY
Huafeng Li,
Kunming University of Science and Technology, China

REVIEWED BY
Yiwen Chen,
Wuhan University, China
Shuanglin Yan,
Nanjing University of Science and Technology, China
Jinting Zhu,
Massey University, New Zealand
Jian Pang,
China University of Petroleum, China

*CORRESPONDENCE
Xiaofeng Wang,
xfwang828@126.com

SPECIALTY SECTION
This article was submitted to Radiation Detectors and Imaging, a section of the journal Frontiers in Physics

RECEIVED 04 December 2022
ACCEPTED 29 December 2022
PUBLISHED 12 January 2023

CITATION
Wang X, Sun J, Qin H, Yuan Y, Yu J, Su Y and Sun Z (2023), Accurate unsupervised monocular depth estimation for ill-posed region.
Front. Phys. 10:1115764.
doi: 10.3389/fphy.2022.1115764

Accurate unsupervised monocular depth estimation for ill-posed region

Xiaofeng Wang[1]*, Jiameng Sun[2], Hao Qin[2], Yuxing Yuan[1], Jun Yu[2], Yingying Su[2] and Zhiheng Sun[2]

[1]College of Mathematical and Physical Sciences, Chongqing University of Science and Technology, Chongqing, China, [2]College of Electrical Engineering, Chongqing University of Science and Technology, Chongqing, China

Unsupervised monocular depth estimation is challenging in ill-posed regions, such as weak texture scenes, projection occlusion, and redundant error of detail information, etc. In this paper, in order to tackle these problems, an improved unsupervised monocular depth estimation method for the ill-posed region is proposed through cascading training depth estimation network and pose estimation network by loss function. Firstly, for the depth estimation network, a feature extraction network using asymmetric convolution is designed instead of traditional convolution, which strengthens the extraction of the feature information and improves the accuracy of the weak texture scenes. Meanwhile, a feature extraction network integrating multi-scale receptive fields with the structure of different scale convolution and dilated convolution stack is designed to increase the underlying receptive field of the depth estimation network, which strengthens the fusion ability of the network for multi-scale detail information, and improves the integrity of the model output details. Secondly, a pose estimation network using an attention mechanism is presented to strengthen the pose detail information of keyframes and suppress redundant errors of the pose information of non-keyframes. Finally, a loss function with minimum reprojection error is adopted to alleviate the occlusion problem of the projection process between adjacent pixels and enhance the quality of the output depth images of the model. The experiments demonstrate that our method achieves state-of-the-art performance on KITTI monocular datasets.

KEYWORDS

unsupervised monocular depth estimation, asymmetric convolution, multi-scale receptive field, attention mechanism, ill-posed regions

1 Introduction

As an important research focus in the field of computer vision, monocular depth estimation aims to explore the mapping relationship between image and depth, and predict the depth information from a single image. Monocular depth estimation plays an important role in visual tasks, especially in intelligent fields such as autonomous driving, 3D map construction, AR (Augmented Reality) synthesis, etc.

At present, the mainstream way of monocular depth estimation task is to train the deep neural network by using a large number of marked real depth images as the training set, so as to obtain the depth value of the corresponding pixel from the image. In this way, deep neural networks are used to generate high-quality depth images with different optimization strategies [1–4]. However, supervised depth estimation methods need to collect a large amount of real-depth information data and require an immense amount of computing time in the training

process, which greatly increases the difficulty and complexity of the algorithm. Comparatively speaking, unsupervised monocular depth estimation only requires monocular video sequences or stereo image pairs to realize the depth information estimation of each pixel of a single image [5–7]. In recent years, unsupervised monocular depth estimation have been favored by researchers [8–10]. Among them, Zhou [10] innovatively proposes an unsupervised training framework which cascades the depth estimation network and the pose estimation network through the loss function to predict the depth information of the image, improving the accuracy of model estimation and becoming one of the most dominant frameworks in current unsupervised monocular depth estimation.

However, current unsupervised monocular depth estimation studies, including Zhou's method, still face great challenges in dealing with ill-posed regions problems, such as weak texture scenes, occlusion of pixel projections, and lack of detailed information in depth images, etc. As a result, the depth information obtained by the model cannot fully reflect the image-depth mapping relationship. To solve these problems, we propose an improved unsupervised monocular depth estimation which included a depth estimation network, pose estimation network, and the loss function. Firstly, in the depth estimation network, asymmetric convolution structure and multi-scale field structure are proposed to enhance the feature extraction capability of the network, to alleviate the influence of weak texture scenes. Secondly, in the pose estimation network, the redundant information of pose estimation of adjacent image frames is reduced by the attention mechanism structure. Finally, the minimum reprojection error is introduced into the loss function to reduce the influence of occluded pixels and inter-frame motion which results in out-of-bounds regions on depth information prediction during pixel projection. By improving the depth estimation network, pose estimation network, and loss function, the accuracy of the unsupervised monocular depth estimation model for depth information is improved, and the robustness and generalization performance of the model is enhanced.

The main contributions of our works are as follows:

- We propose an unsupervised monocular depth estimation method improved for ill-posed regions by training a depth estimation network and a pose estimation network in cascade with loss functions.
- We improve the unsupervised depth estimation network by using asymmetric convolution, multiscale perceptual field structure, SE structure and minimum reprojection error in ill-posed regions, such as weak texture scenes, pixel projection occlusion, lack of detailed information in depth images, and so on.
- Our approach demonstrate state-of-the-art performance at KITTI monocular datasets.

2 Approach

At present, the unsupervised monocular depth estimation model takes video sequences as input and constructs an unsupervised learning framework for monocular depth and camera pose estimation based on unstructured video sequences. Specifically, an end-to-end learning method is used to jointly train a depth estimation network and a pose estimation network in an encoder-decoder manner, so as to obtain the depth information in a single frame of a video sequence in an unsupervised manner [11].

However, current unsupervised monocular depth estimation algorithms still have limitations when dealing with ill-posed regions, such as weak texture scenes, occlusion of pixel projection, detail information lack of depth images, and redundant errors of continuous image frames for pose information.

In order to further improve the unsupervised monocular depth estimation model and cope with the above complex scenes, this paper improves the unsupervised monocular depth estimation model, which consists of depth estimation network, pose estimation network, and the loss function. We predict the depth information and pose information of 2D images by cascading the depth estimation network and pose estimation network, then we take the pixel error between the reconstructed image and the input image as the supervised signal of the whole network to achieve the depth estimation of unsupervised monocular estimated images. Firstly, for the depth estimation network, inspired by Ding [11], the AC (Asymmetric Convolution) is designed to extract the features of the input image from vertical, horizontal, and overall directions, so as to alleviate the influence of weak texture scenes. Through RFB (Receptive Field Block) which is a multi-scale receptive field structure [12], the ability to obtain all and local information is enhanced in the receptive field area of different scales of the network. Secondly, for the pose estimation network, SE(Squeeze-and-Excitation) structure [13] is introduced to reduce the error region of pose estimation. Finally, for the loss function, the concept of minimizing reprojection error is introduced to reduce the impact of pixel projection occlusion in depth information estimation.

The overall structure of improved unsupervised monocular depth estimation network is shown in Figure 1. Firstly, multi-scale feature maps which is equivalent to 1/2, 1/4, 1/8, 1/16 resolution of the input image frame are generated in the improved depth estimation network, and then these features are mapped to the depth decoder with parameter sharing, and the estimated depth is restored to the same size as the resolution of the input image frame through the upsampling structure. Secondly, for the improved pose estimation network, the relative pose of 6 degrees of freedom which includes displacement with 3 degrees of freedom and spatial rotation with 3 degrees of freedom is generated by the pose estimation network. Finally, the depth information and pose information obtained by the improved depth estimation network and pose estimation network are jointly trained using the loss function.

2.1 The depth estimation network optimization

At present, most unsupervised monocular depth estimation algorithms cannot effectively deal with weak texture scenes and miss detailed information of the predicted depth image. In order to solve this problem, asymmetric convolution and multi-scale receptive field RFB are used in the depth estimation network to enhance the recognition of weak texture scenes and strengthen the acquisition of detailed information. The depth estimation network is improved accordingly.

2.1.1 Improved ACResNet50 depth estimation network

Weak texture regions are not distinct and significant features, which are prone to semantic ambiguity and lead to wrong depth

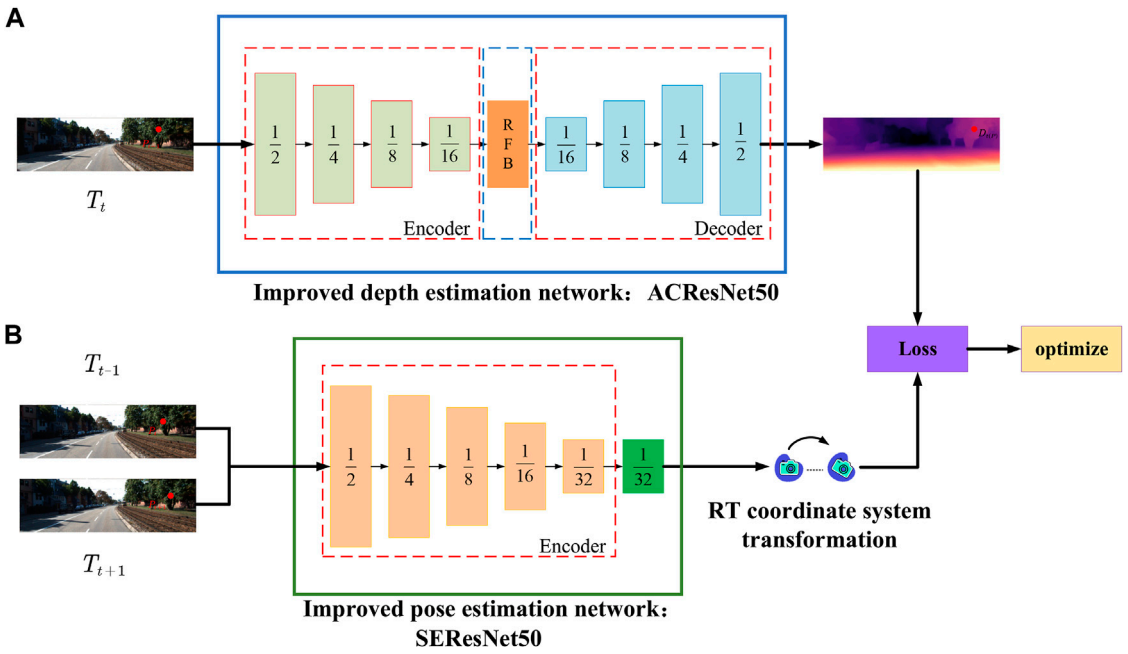

FIGURE 1
Structure diagram of unsupervised monocular depth estimation model: **(A)** Improved depth estimation network ACResNet50; **(B)** Improved pose estimation network SEResNet50.

estimation, so we deal with this problem in this paper. Our improved ACResNet50 depth estimation network is shown in Figure 2.

Firstly, in order to effectively mitigate the impact of weak texture scenes on the accuracy of depth estimation information, the traditional convolution method is replaced by AC, and each traditional convolution of ResNet50 network is replaced to strengthen its feature extraction ability, and the ACResNet50 network structure is formed. Secondly, in order to solve the problem of missing details, the RFB structure is connected to the last structure of ACResNet50. Based on the convolution of different sizes, the dilated convolution is added and its expansion rate is adjusted to ensure the network receptive field, so as to achieve the acquisition of high-resolution features. The fusion of global feature information and local feature information is strengthened. Finally, the deconvolution structure is used to restore the size of the output feature map to the size of the original feature image, and the prediction function of the entire network depth information is realized.

2.1.2 Asymmetric convolution

At present, the unsupervised monocular depth estimation network performs poorly on weak texture scenes. Most unsupervised monocular depth estimation networks use the ResNet50 network as the feature extraction backbone network in the encoding process of the image and extract the feature information of the image by feature superposition and refinement. Although the residual structure of ResNet50 can extract the feature information of the image to a certain extent, it is far from sufficient for the task of unsupervised monocular depth estimation that requires more accurate depth information. At the same time, the continuous superposition of the ResNet50 network and the deepening of the number of network layers will also lead to many problems, such as too many network parameters, difficult training, and the degradation of the whole network.

In order to obtain more feature information of the input image and alleviate the influence of weak texture scenes on unsupervised monocular depth estimation tasks, inspired by ACNet research, the traditional convolution method is improved, and we propose a novel depth estimation network based on ACNet. The feature extraction of the input image is carried out from the vertical, horizontal, and overall directions, which strengthens the feature information extraction ability of the feature extraction network and alleviates the influence of weak texture scenes on the depth information.

Figure 3 is the operation process of asymmetric convolution. The ACNet network in the Figure 3 can be divided into two stages, training and test reasoning, with Figure 3A indicating the training stage and Figure 3B indicating the test reasoning stage. Firstly, we set up three parallel convolution kernels with sizes 1×3, 3×1, and 3×3 respectively, 1×3 and 3×1 convolution kernels facilitate the extraction of edge information of weak texture regions and other regions to identify weak texture regions with other regions, and then joint 3×3 convolution to extract contextual features of weak texture regions to improve the accuracy of weak texture depth estimation. Secondly, the input image is processed by these three parallel convolution kernels respectively, so that the extracted feature information has the characteristics of horizontal, vertical, and overall directions, then the three kinds of feature information can be stacked and output. Finally, the traditional convolutions in the network are replaced with non-traditional convolutions to form the improved feature map extraction network on ResNet50.

2.1.3 Multiscale receptive fields

The lack of details in depth maps has always been a difficulty for unsupervised monocular depth estimation. The reason is that in the

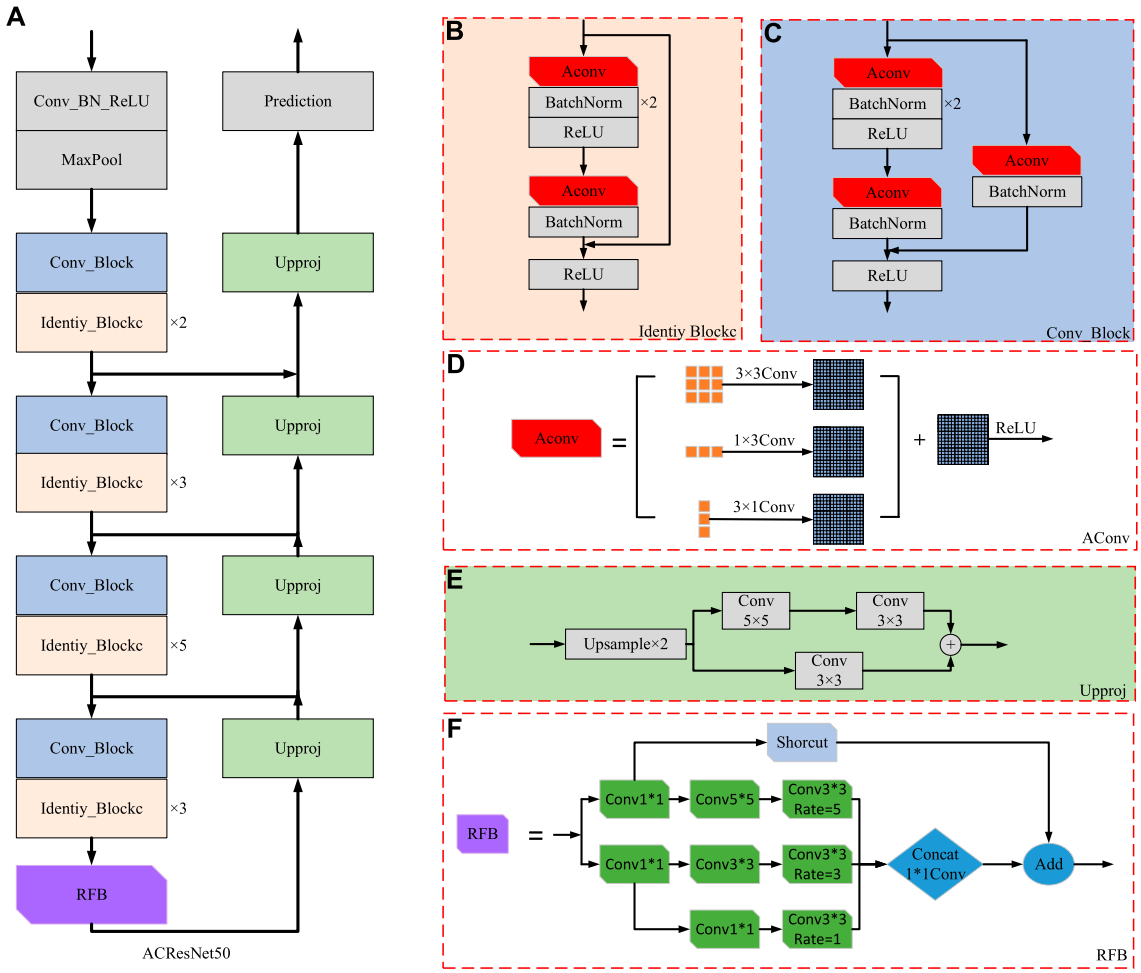

FIGURE 2
Depth estimation network model: **(A)** ACResNet50 structure; **(B)** Identiy_Block structure; **(C)** Conv_Block structure; **(D)** AConv structure; **(E)** Upproj structure; **(F)** RFB structure.

theory of deep convolutional neural networks, the perceptual field of the network gradually increases with the number of layers of the network, Zhou [14] founds that the network's ability of detail acquisition in the receptive field is reduced in deeper networks, leading to poor network learning. Moreover, in the traditional convolution process, convolution is used to continuously stack down sampling to extract abstract information, but continuous downsampling will lead to the loss of image details and local information. Zhao [15] points out that the fusion of global and different scale context information in semantic segmentation is beneficial to alleviate the loss of detail information and preserve the spatial structure of the image. Therefore, RFB is adopted to solve this problem in that the receptive field decreases in the unsupervised monocular depth estimation model, which leads to unsatisfactory context information fusion and missing details of the estimated depth map.

The RFB can achieve the acquisition of high-resolution features without repeated down-sampling and enhance the ability of network feature extraction and fusion [12]. At the same time, different receptive fields are obtained by adjusting different expansion rates of dilated convolution, so as to enhance the variability of network receptive field region size. By stacking in this way, the ability of interfusion between feature information at different scales of the network is enhanced, and the acquisition of full and local detail information is strengthened. The multiscale receptive field RFB structure is shown in Figure 4.

In this paper, a multi-scale receptive field RFB structure is added after the last convolutional block of the ACResNet50 feature extraction network. Firstly, in the multi-branch convolution layer, convolution kernels of 1×1, 3×3 and 5×5 sizes are used to ensure the performance of the network to deal with scale changes and improve the multi-scale feature extraction ability of the model. Secondly, on the dilated convolution, in order to ensure the consistency of the scale of the multi-branch convolution layer and the expansion rate of the dilated convolution, by connecting cavity convolution with expansion rates of 1, 3 and 5, respectively to convolution of different scales, we enhance the receptive field of the network and improve the acquisition ability of high-resolution feature maps and context information. Finally, the image feature information of different scales is fused by stacking the features to generate a receptive field spatial array as the input of deconvolution through 1×1 convolution.

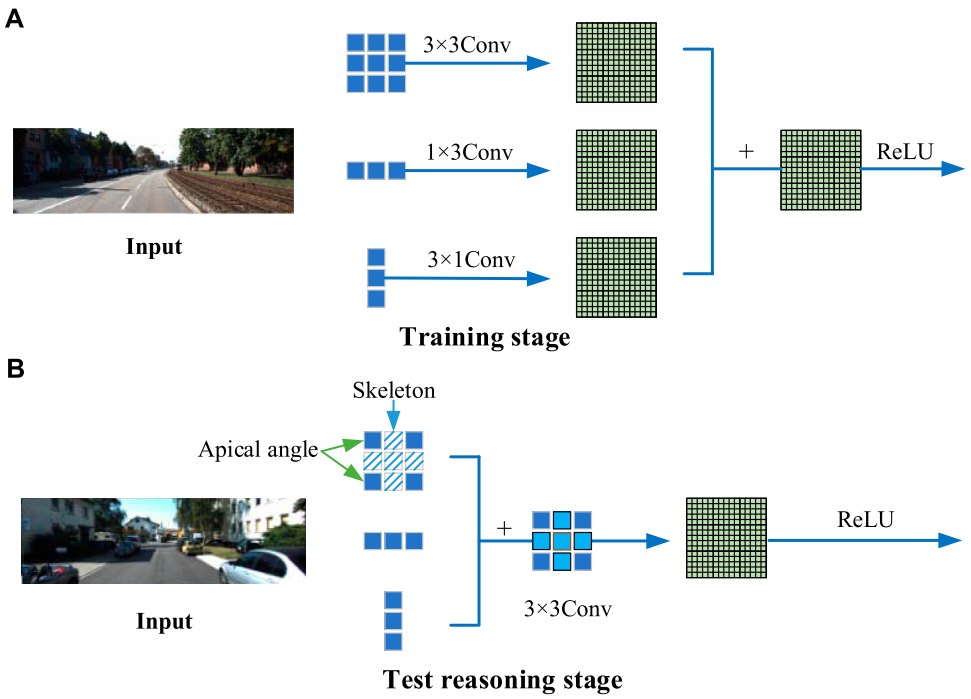

FIGURE 3
Asymmetric convolution structure:**(A)** The ACNet structure in the training stage; **(B)** The network structure in the test reasoning stage.

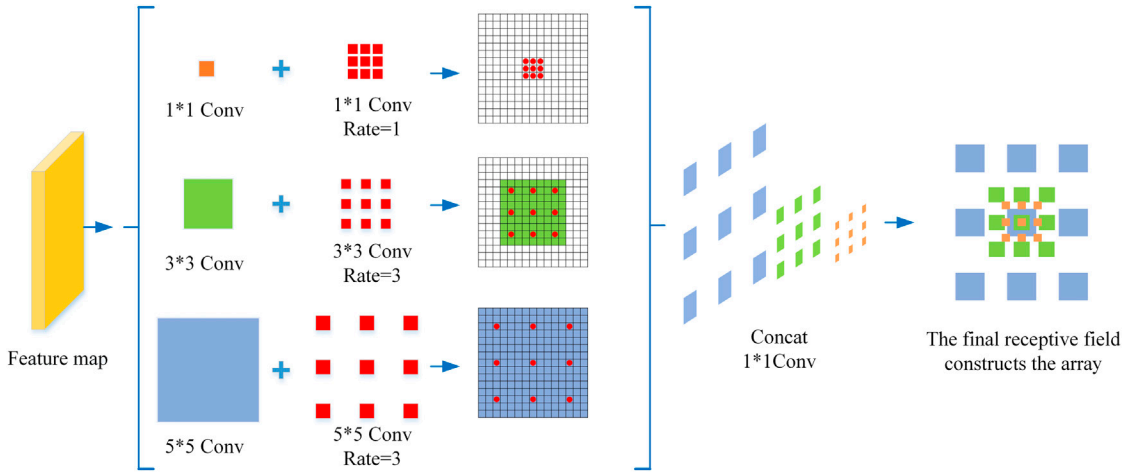

FIGURE 4
Multi-scale receptive field RFB structure.

2.2 Pose estimation network based on SEResNet50

The pose estimation network is crucial for accurately predicting depth information. However, in the design process of the pose estimation network, most unsupervised monocular depth estimation models directly use the pose information of consecutive image frames for model prediction, ignoring the redundant error of pose information, which leads to the reduction of the accuracy of the model prediction depth information.

In order to reduce the large redundant errors in pose estimation, we design the SE attention mechanism structure based on ResNet50 in the pose estimation network, which can focus on the important pose information of the image frame, suppress the unimportant pose information of the image frame, and reduce the large error redundancy. The improved pose estimation network is shown in Figure 5.

2.2.1 SE attention mechanism

For the pose estimation network, its task is to accurately predict the camera motion trajectory between adjacent frames in the video

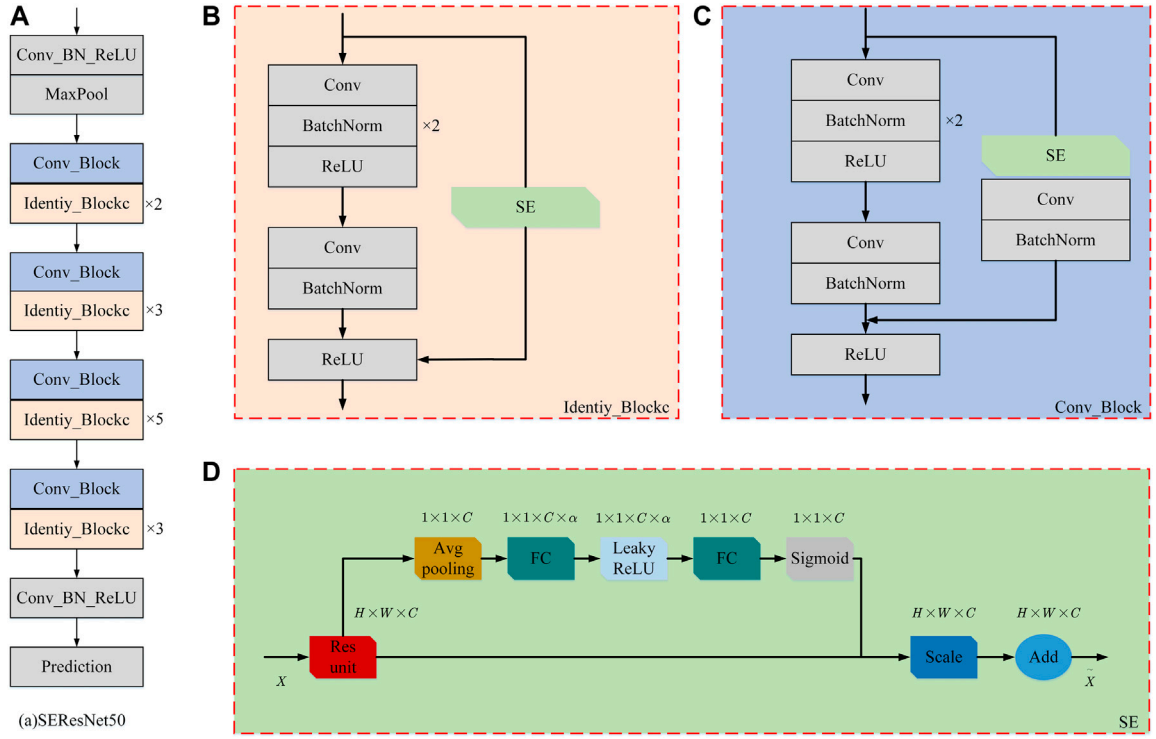

FIGURE 5
Improvement of pose estimation network model: **(A)** SEResNet50 structure; **(B)** Identiy_Block structure; **(C)** Conv_Block structure; **(D)** SE structure.

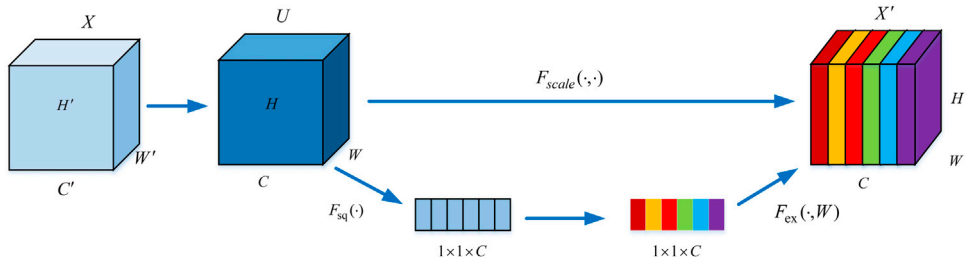

FIGURE 6
SE channel attention mechanism.

sequence, so as to obtain the rotation matrix and translation matrix. Then, the image is reconstructed by combining the internal parameter matrix of the camera and the depth information predicted by the depth estimation network. However, in the pose estimation network, the camera pose motion estimation in the image between two highly adjacent frames is highly approximate. If the network trains all the pose information of the video sequence frames and predicts the camera pose, it will not only increase the amount of information processed by the network but also lead to an increase in redundancy error in the pose estimation.

SE structure focuses on exploring the relationship between different channels in feature information, and this exploration method has a good performance in balancing the importance of feature channels and learning global feature information [13]. In the pose estimation network, the SE structure can be used to pay more attention to the important pose information in the continuous image frames of the video sequence, suppress the unimportant pose information, and effectively enhance the network's prediction of the camera pose motion trajectory between image frames, and improve the ability of pose estimation.

Figure 6 shows the attention mechanism structure of SE channel. Firstly, given a feature input X, its height, width, number of channels, and the dimension are H', W', C', and $H' \times W' \times C'$, respectively. After a series of transformations such as convolution, a feature U of size $H \times W \times C$ is obtained. Secondly, the squeeze operation $F_{sq}(\cdot)$ is carried out, so that the feature U is squeezed along the spatial dimension. Further, each two-dimensional characteristic channel is turned into a real number, and a feature with the same dimension and channel number is output, whose size is $1 \times 1 \times C$. Thirdly, the excitation operation $F_{ex}(\cdot, W)$ is used to generate a weight for each feature channel, where W represents the correlation between feature channels. Finally, by doing the scale operation $F_{scale}(\cdot,\cdot)$, the weight

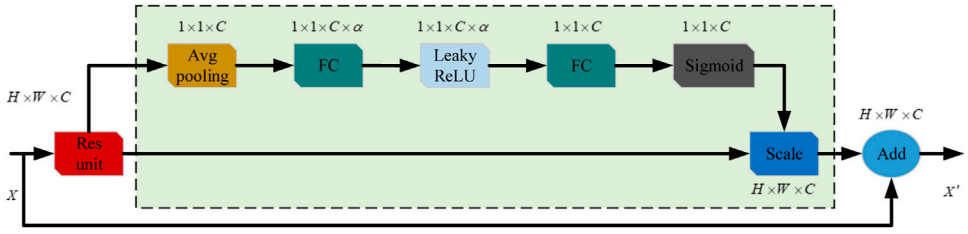

FIGURE 7
SE+ Residual blocks structure.

output by the excitation operation is weighted to the feature U of the previous layer channel by channel through multiplication, and the feature X' with attention mechanism is obtained with size $H \times W \times C$. Through the above operations, we retain important feature information and strengthen learning ability of global feature information.

2.2.2 Residual network with attention mechanism

There are multiple block structures in the original ResNet50 feature extraction network, and each block realizes the extraction of image features by stacking each other. The original ResNet50 backbone feature extraction network does not consider the relationship between different channels in the feature information. Such a way will lead to the lack of ability to distinguish the main and secondary channel feature information, resulting in a weak performance in global feature information extraction ability.

However, the SE channel attention mechanism makes full use of the weights of different feature channel information importance to enhance the information acquisition of important feature channel. In this paper, the attention mechanism is introduced into the backbone feature extraction network to strengthen the extraction performance of global feature information. Its improved residual network with attention mechanism is shown in Figure 7.

This paper designs the channel attention mechanism SE in each block structure of the ResNet50 feature extraction network. Firstly, the feature map of height H, width W, number of channels C and size $H \times W \times C$ by ResNet50 is output a feature map of size $1 \times 1 \times C$ by global average pooling. Secondly, the feature map of $1 \times 1 \times C$ is input to the first fully connected layer with ReLU as the activation function and the output is $1 \times 1 \times C \times \alpha$, where the number of neurons is $C \times \alpha$ and the scaling parameter is α , which aims to reduce the channel reduction calculation. It is input to the second fully connected layer with sigmoid as the activation function, whose output is $1 \times 1 \times C$ and the number of neurons is C to complete the acquisition of the weight of the attention mechanism of different channel feature information. Then, the obtained weights are applied to the $H \times W \times C$ feature information of ResNet50 output through the multiplication operation to obtain the feature channels with weights. Finally, the feature output of the previous layer and the weighted feature channel are superimposed to obtain the final feature output. An improved SEResNet50 residual block with attention mechanism is formed. This structure can effectively enhance the performance of the network in extracting feature information of important adjacent frame image pose changes and reduce the redundancy error of the pose estimation network.

2.3 Design of the loss function

In the design of the loss function, since the whole unsupervised monocular depth estimation network consists of two parts: the depth estimation network and the pose estimation network, which are used together to predict the depth of a pixel. Therefore, the constraint term of the loss function is derived from the pixel difference between the reconstructed image and the input image after information predicted by the depth estimation network and the pose estimation network. In the inference of the loss function, let the three adjacent frames of images at time t be I_t, I_{t-1}, and I_{t+1}.We call I_t the target image and the other two I_{t-1} and I_{t+1} the source images. Firstly, the depth $\hat{D}_t(p_t)$ of each pixel p_t in the target view I_t is obtained through the depth estimation network, and then (I_t, I_{t-1}) and (I_t, I_{t+1}) are fed into the pose estimation network as a group to obtain the camera motion $\hat{T}_{t\to t-1}$ and $\hat{T}_{t\to t+1}$ between neighboring pixels respectively. In this way, the depth information and pose information of the color image are obtained.

In the process of image reconstruction, each pixel p_t in the target view I_t is projected onto the source image $I_s \in (I_{t+1}, I_{t-1})$ at pixel s according to the predicted depth information $\hat{D}_t(p_t)$ and camera pose $\hat{T}_{t\to t-1}, \hat{T}_{t\to t+1}$. Bilinear interpolation is then used to obtain p_t which is the value of the distorted image. The differentiable image warping process is shown in Figure 8.

For the flush coordinate p_t of a pixel in the target frame, then the projection coordinate of p_t corresponding to the p_s of the source frame can be obtained as follows:

$$p_s \sim K\hat{T}_{t\to s}\hat{D}_t(p_t)K^{-1}p_t, \tag{1}$$

where $\hat{T}_{t\to s}$ is the camera motion pose from frame t to s, $\hat{D}_t(p_t)$ is the depth value of pixel p_t in frame t, and K is the camera internal reference matrix.

In this case, let the target image I_t of the reconstructed frame, the source image I_s as the frame used to reconstruct I_t, and the reconstructed image $\hat{I}_s$. Let $< I_1, \ldots, I_N >$ be a training image sequence, where one of the frames is denoted as the target image I_t. I_s is the source image sequence denoted as $I_s (1 \le s \le N, s \ne t)$. $||$ measures the absolute error. Then the loss function L_1 is expressed as follows:

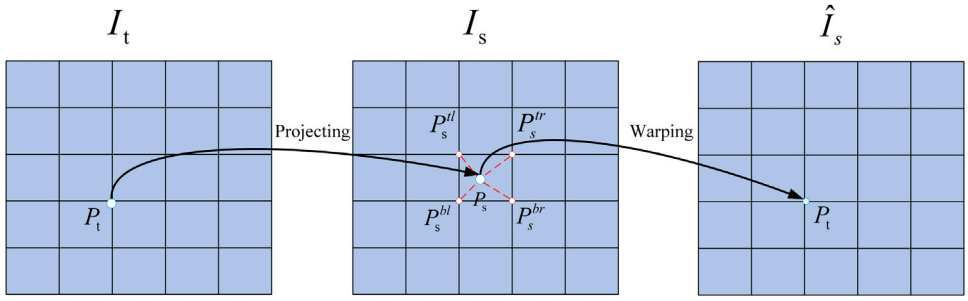

FIGURE 8
Differentiable image warping process.

$$L_1 = \sum_s \sum_p \left| I_t(p) - \hat{I}_s(p) \right|. \tag{2}$$

Since the premise assumptions of invisible change and static scene need to be satisfied in network construction, if one of the assumptions is not met, the gradient will be destroyed and the inhibition of training will occur. In response to these factors, in order to improve the robustness of the network, the output confidence weight $\hat{E}_S(p)$ for each target source pair is given during the cascaded training of the depth estimation network and the pose estimation network. After weighting the loss function (2), the loss function L_2 is expressed as:

$$L_2 = \sum_s \sum_p \hat{E}_s(p) \left| I_t(p) - \hat{I}_s(p) \right|. \tag{3}$$

In the original algorithm, the ill-posed region is solved by adding a smoothing constraint when obtaining the depth map, and the depth of each pixel is solved by global optimization. However, this method is to average the reprojection error of multi-source images, which may lead to problems of pixels which are visible in the target image and invisible in the source image. If the network predicts the correct depth of a pixel, then the corresponding color in the blocked source image has a high probability of mismatch with the target, resulting in a high photometric error. There are two reasons for this problem. One is pixels which are on the edge of the image and are out of view due to motion between frames. The other is the occluded pixels.

In this paper, we use the concept of minimum reprojection error to deal with the problem of out-of-bounds caused by occluded pixels and inter-frame motion. At each pixel, the photometric error of all source images is no longer averaged, but simply the minimum value is used, which can effectively alleviate the pixels that are visible in the target image and invisible in the source image in the process of pixel projection, and solve the occlusion problem caused by pixel projection. Therefore, the calculation process of the minimum reprojection loss function L_p is as follows:

$$L_p = \sum_{t'} pe(I_t, I_{t' \to t}), \tag{4}$$

$$pe(I_a, I_b) = \frac{\alpha}{2}(1 - SSIM(I_a, I_b)) + (1 - \alpha)\|I_a - I_b\|_1. \tag{5}$$

Among them, SSIM (Structural Similarity Index Measurement) is the structural similarity index, I_a and I_b are the adjacent frame images, t represents the time of each frame image. As known relative pose at time t', the source image $I_{t'}$ is the second frame in the stereo pair to I_t, and α is set to .85 to make its edge perception smooth.

Finally, the minimum reprojection error constraint is introduced into the overall loss function to reduce the impact of pixel occlusion on the model during pixel projection and ensure the accuracy of the model in predicting depth information. The final loss function of the model L_{final} is as follows:

$$L_{final} = \sum_l L_1^l + \lambda_p L_p^l + \lambda_e \sum_s L_{\text{reg}}\left(\hat{E}_s^l\right). \tag{6}$$

λ_p and λ_e are the weight value of minimizes the reprojection error and the weight value normalized by the target source on the output confidence. We empirically take the values are .65 and .35, respectively. L_{reg} denotes the regularization term [16], and l represents different image scales, respectively.

3 Experiments

In our experiment, video images of real scenes are utilized as training data set and test data set, such as urban areas and highways in KITTI data set. In order to ensure the consistency of the experiment, the image resolution is uniformly cropped to a size of 640 × 192. The common methods of data enhancement such as rotation and flip are also used to expand the data. The SGD (Stochastic Gradient Descent) algorithm is used to optimize the model parameters. The training iteration epochs of the whole network is set to 200. The initial learning rate is set to .001 and dynamic attenuation is adopted. Image acceleration is CUDA11.2.0/CUDNN8.2.1.

3.1 The ablation experiment

To verify the reliability of the proposed scheme, we validated the proposed scheme on the KITTI dataset and performed ablation experiments and compared the proposed method in this paper with the scheme of Zhou [10], and the experimental scheme and results are shown in Table 1.

3.2 The depth estimation network

3.2.1 Verification of asymmetric convolution structure

In order to verify AC, we conduct comparative experiments between ACResNet50 in this paper and Zhou's method.

From the quantitative and qualitative analysis of the relevant evaluation indicators in Table 1 and Figure 9, our method works

TABLE 1 Ablation experimental design protocol and comparison of experimental results.

Category of schemes					Error metric			Accuracy metric		
Zhou [10]	ACResNet50	RFB	SEResNet 50	L_p	Abs rel	Rmse	Rmse log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
√	×	×	×	×	0.183	6.709	0.27	0.734	0.902	0.959
√	√	×	×	×	0.169	6.391	0.262	0.740	0.91	0.963
√	√	√	×	×	0.164	6.249	0.258	0.758	0.915	0.965
√	√	√	√	×	0.162	6.211	0.246	0.773	0.918	0.968
√	√	√	√	√	0.161	6.032	0.235	0.781	0.922	0.970

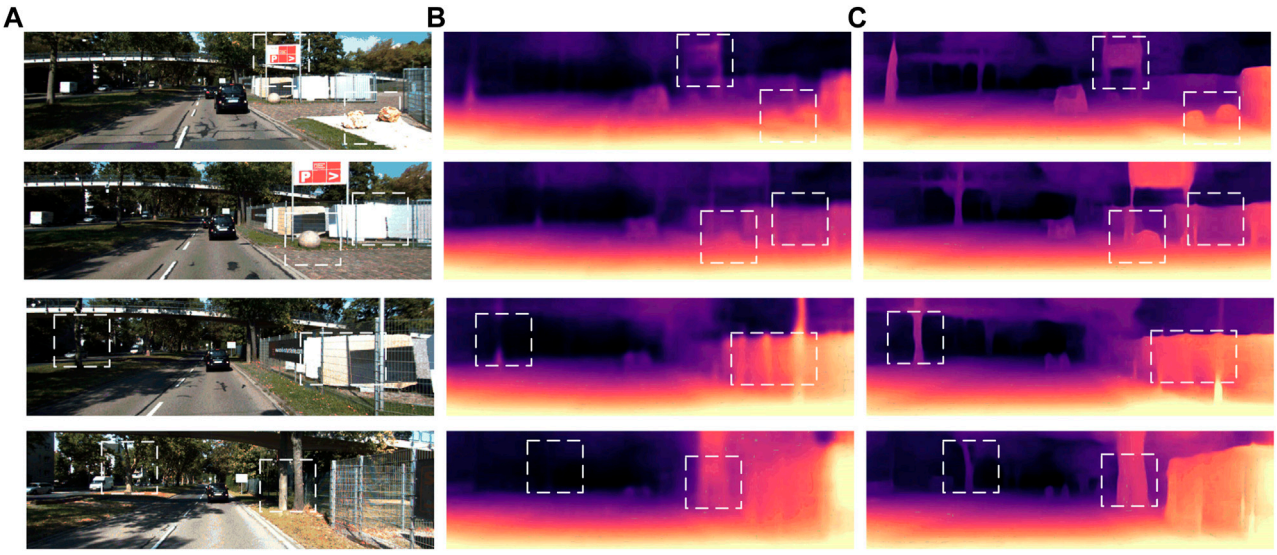

FIGURE 9
Comparative experimental results of asymmetric convolution module visualization on KITTI dataset: **(A)** represents the color image input by the model; **(B)** represents the depth map result predicted by Zhou et al.; **(C)** The ACResNet50 prediction of asymmetric convolution to the depth map results.

better than Zhou's method for weak texture scenes in the ill-posed regions of the billboard in row 2 and the columnar objects in rows 2 to 3. Only using the asymmetric convolution structure designed to replace the traditional convolution structure has a certain improvement in the accuracy of the estimated depth value of the network. The experimental results show that the improved asymmetric convolution can effectively enhance the ability of the network to obtain feature information for the color two-dimensional image, strengthen the feature extraction of the input image, and make the unsupervised monocular depth estimation network output depth images with rich textures and clear edges.

3.2.2 Validation of ACResNet50+ RFB structure

In order to verify the RFB structure, the ACResNet50 + RFB is compared with Zhou [10].

The relevant quantity and quality evaluation metrics are analyzed in Table 1 and Figure 10. In this paper, RFB is introduced into the last module of the ACResNet50 network, so that the model can obtain the context information of image features at different scales. The obtained feature information is more continuous and the detail information is more complete, which ensures the continuity and integrity of the spatial structure of the output depth image of the network. In Figure 10, our method is able to retain more detailed information of vehicle contours, which is significantly better than Zhou's method. Experimental results show that the proposed multi-scale receptive field enhanced RFB structure outperforms Zhou's algorithm in depth map detail information and spatial structure presentation. It can effectively avoid the lack of details in the unsupervised monocular image depth estimation task, strengthen the control of the model for detailed information. At the same time, it can further obtain multi-scale information and rich context information in two-dimensional color images, and improve the overall prediction accuracy and generalization performance of the model. The results show that the method can effectively alleviate the redundancy error problem of detail information in the ill-posed regions.

3.3 Pose estimation network

In order to verify the actual effect of the pose estimation network SEResNet50 embedded with the attention mechanism designed in this paper, the method in this paper is compared with Zhou [10].

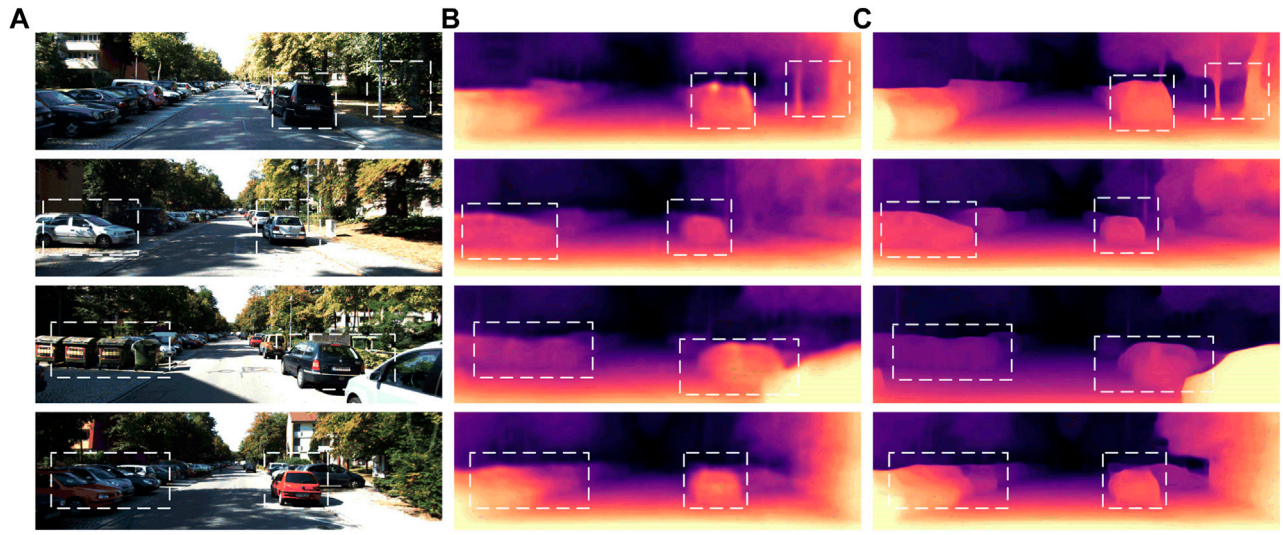

FIGURE 10
Comparative experimental results of multi-scale receptive field RFB visualization on KITTI dataset: **(A)** The color images input by the model; **(B)** The depth map results predicted by Zhou et al.; **(C)** The depth map results predicted by ACResNet50+ RFB structure.

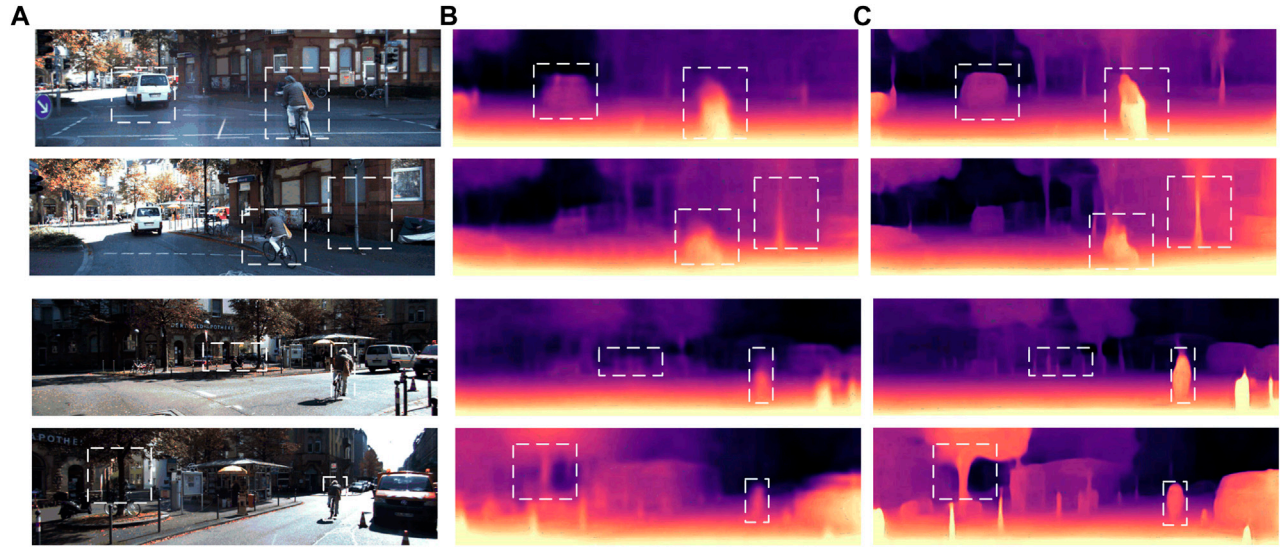

FIGURE 11
Comparative experimental results of attention mechanism SE module visualization on KITTI dataset: **(A)** The color image input by the model; **(B)** The depth map results predicted by Zhou et al.; **(C)** The depth map predicted by SEResNet50 embedded attention mechanism in the pose estimation network.

The relevant evaluation indicators in Table 1 and Figure 11 are analyzed quantitatively and qualitatively. In this paper, the attention mechanism SE structure is designed to reduce the redundant error caused by using the pose information of consecutive frames to predict the pose information of the next frame in the pose estimation process. The attention mechanism SE can pay attention to the important information in a single frame and suppress the unimportant information, so as to effectively reduce the redundant error generation and improve the overall prediction accuracy of the model. From Figure 11, we can find that our method works well when targeting the projected occlusion region of bicycle pedestrians and car outline. The experimental results show that the attention mechanism SE structure designed in this paper can reduce the redundant error of camera pose estimation in the pose estimation network. In terms of the accuracy of predicting the depth value, the three indicators have a corresponding improvement, where $\delta < 1.25$, it is an obvious improvement over Zhou [10], and the output depth map is of high quality. It shows that the pose estimation network designed in this paper can effectively estimate the motion pose of the camera accurately, and it is a good contribution to the whole unsupervised monocular depth estimation network to predict depth information.

At the same time, in order to further verify the absolute trajectory error estimated as the pose information, the prediction results are

TABLE 2 Absolute trajectory error for validating positional estimation on KITTI test set.

Methods	Seq.9	Seq.10
ORB-SLAM (full) [17]	0.014 ± 0.008	0.012 ± 0.011
ORB-SLAM (short) [17]	0.064 ± 0.141	0.064 ± 0.130
Zhou [10]	0.021 ± 0.017	0.020 ± 0.015
Our method	0.019 ± 0.015	0.018 ± 0.016

tested through the pose estimation test data Seq.9 and Seq.10 provided by the KITTI dataset official website, as shown in Table 2.

As can be seen from Table 2, after designing the attention mechanism in the pose estimation network, the error of pose estimation on the KITTI test set is smaller than that of ORB-SLAM (short) and Zhou's method, but larger than that of ORB-SLAM (full). Therefore, the attention mechanism used in the pose estimation network can effectively reduce the redundant error caused by the superposition of consecutive multi-frame image information and improve the robustness of the model.

3.4 Minimum reprojection error loss function

In order to verify the experimental effect of introducing the minimum reprojection error loss function. The model introduced with the minimum reprojection error loss function designed in this paper is compared with the method of Zhou [10].

The relevant evaluation indicators in Table 1 and Figure 12 are analyzed quantitatively and qualitatively. In this paper, a constraint term of minimum reprojection error is added to the loss function, which is beneficial for the prediction of depth information, and can effectively improve the occlusion problem in the projection process of adjacent pixels.

Experimental results show that after using the minimum reprojection error as a constraint term, each error index is reduced accordingly. It improves the problem of occlusion during the projection of adjacent pixels and enhances the prediction accuracy of depth information of the model. At the same time, the robustness and generalization performance of the model are improved.

3.5 KITTI contrast experiment

At the same time, in order to verify the effectiveness and generalization of the proposed method, we make qualitative and quantitative comparison analysis with the research algorithms in related fields. In order to verify the effectiveness of the method in this paper, the comparative experiments are based on the KITTI dataset, verify the generalization of the method in this paper, the cityscapes dataset is used, but the error of the model increases slightly when dealing with data sets other than KITTI.

In Table 3, k is the KITTI dataset, CS is the Cityscapes dataset, and supervision (Y, N) indicates whether it is an unsupervised and supervised monocular depth estimation task. The relevant evaluation indicators in Table 3 and Figure 13 are analyzed quantitatively and qualitatively. The algorithm designed in this paper is .022, .677, and .035 lower than Zhou in AbsRel (Absolute Relative error), RMS (Root Mean Square error), and LogRMS (Log Root Mean Square error), respectively. In the three depths value accuracy evaluation indicators of $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$, it is .047, .020, and .013 higher, respectively. The accuracy of predicting depth information from monocular color image is also better than that of the algorithm proposed by Zhou [10].

The method designed in this paper has good performance in various evaluation indicators compared with the previous research work. Among them, compared with the supervised method of Eigen [18], Liu [19] and Cao [22], the accuracy of the predicted depth value is greatly improved. Compared with the unsupervised monocular depth estimation proposed by Zhou [10], the three indexes in this paper are increased by .047, .020, .013 respectively, and the error index is reduced accordingly. Compared with the recent work of Yang [21], AdaDepth [23], S2R-DepthNet [24], etc. which studied the unsupervised monocular depth estimation task, the proposed method performs better in all indicators. At the same time, from the depth images predicted by each algorithm in Figure 13, the proposed algorithm has good performance in the texture information, detail information, and spatial structure of the output depth map.

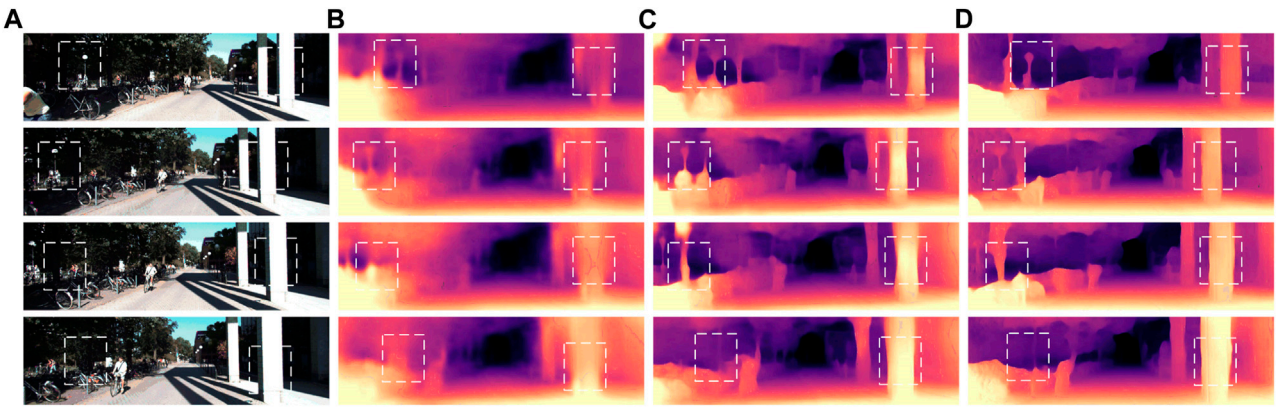

FIGURE 12
Comparative experimental results of the per-pixel minimum reprojection error visualized in the KITTI dataset: **(A)** The color image input by the model; **(B)** The depth map result predicted by Zhou et al.; **(C)** The depth map predicted by the whole network structure after improvement; **(D)** The depth map predicted by the whole model after using the minimum reprojection error loss function.

TABLE 3 Comparison of experimental results with other related research algorithms.

Methods	Supervised	Data	Error			Accuracy, δ		
			AbsRel	RMS	LogRMS	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen [18]	Y	K	0.214	6.307	0.292	0.673	0.884	0.957
Liu [19]	Y	K	0.202	6.471	0.275	0.678	0.895	0.965
Zhou [10]	N	K	0.183	6.709	0.27	0.734	0.902	0.959
UnDeepVO [20]	N	K	0.183	6.57	0.268	—	—	—
Yang [21]	N	K	0.182	6.501	0.267	0.725	0.906	0.963
Cao [22]	Y	K	0.180	6.311	—	0.771	0.917	0.966
AdaDepth [23]	N	K	0.167	5.578	0.237	0.771	0.922	0.971
S2R-DepthNet [24]	N	K	0.165	5.695	0.236	0.781	0.931	0.972
Geonet [25]	N	K	0.164	6.09	0.247	0.765	0.919	0.968
Mahjourian [26]	N	K	0.163	6.22	0.25	0.762	0.916	0.966
LEGO [27]	N	K	0.162	6.276	0.252	—	—	—
Our method	N	K	0.161	6.032	0.235	0.781	0.922	0.972
Our method	N	CS	0.174	6.322	0.259	0.748	0.911	0.964
Our method	N	K + CS	0.168	6.282	0.26	0.731	0.908	0.963

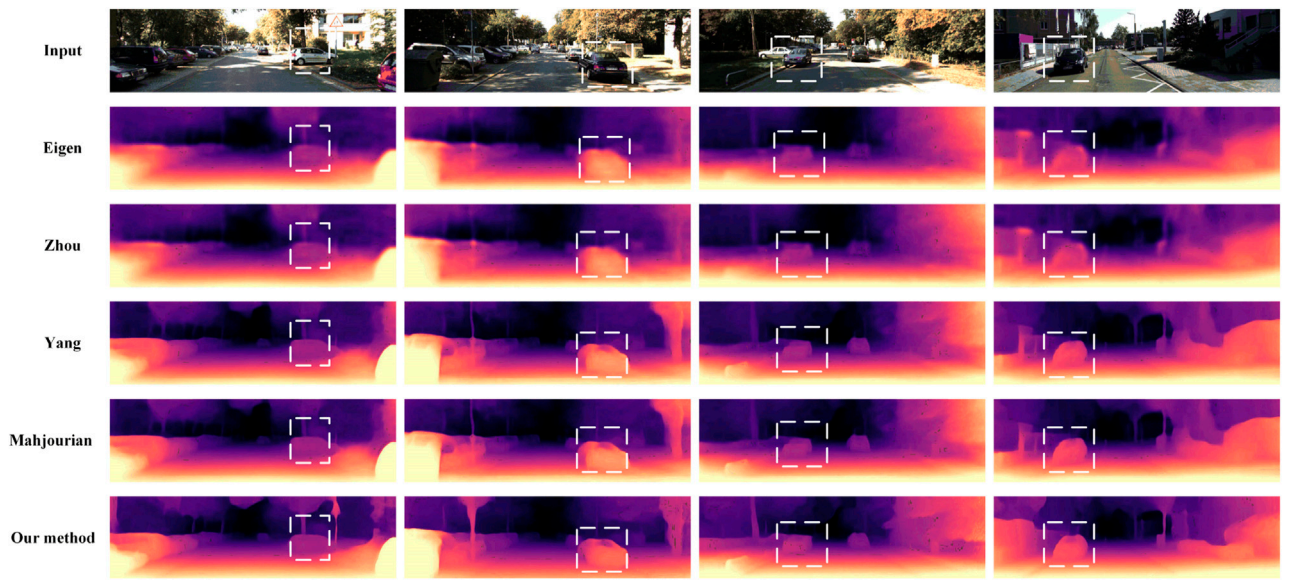

FIGURE 13
Visual comparison experimental results with other related research algorithms on the KITTI dataset: the first row is the input two-dimensional color image; rows 2 to 6 show the depth maps predicted by Eigen, Zhou, Yang, Mahjourian et al. and our method, respectively.

The experimental results show that the improved unsupervised monocular depth estimation algorithm designed in this paper can effectively alleviate the impact of weak texture scenes on the model, solve the lack of detail of the input image, reduce the redundant error of pose information, reduce the occlusion problem in the process of pixel projection, and ensure the prediction accuracy of the unsupervised monocular depth estimation model. From the analysis of the above indicators, the unsupervised monocular depth estimation network has a certain competitive advantage in depth prediction, and can accurately estimate the depth information of images or video frames.

4 Conclusion

Currently, supervised monocular image depth estimation tasks require a large amount of real depth data for training, which greatly

increase the development cost of the model and the difficulty of landing the model. The improved unsupervised monocular depth image estimation task designed in this paper only uses continuous video sequences to complete the depth prediction of each pixel of a single image, which greatly reduces the model development cost and accelerates the model implementation process. It can effectively improve the influence of weak texture scene on depth prediction, reduce the lack of details of the model predicted depth image, and reduce the occlusion problem of the model due to the pixel projection process. Through the improvement of this paper, the prediction accuracy of the unsupervised monocular image depth estimation model on depth information is strengthened, which makes the depth image predicted by the model richer in texture information, clearer in detail information, and more continuous in spatial structure, thus enhancing the structure of the predicted depth image and improving the resolution of the output image. The robustness and generalization performance of the unsupervised monocular depth estimation model are improved.

Although our approach does not require labeling of real depth images as supervised methods do, the framework lacks explicit estimation of scene dynamics in 3D scene understanding. In future work, we would like to explore methods for modeling scene dynamics through motion segmentation to improve the performance of unsupervised monocular depth estimation in dynamic scenes.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

Conceptualization, XW; methodology, XW and JS; software, YS; validation, JY, YS, and ZS; formal analysis, XW; writing—original draft preparation, YY; writing—review and editing, XW; supervision, XW and HQ. All authors have read and agreed to the published version of the manuscript.

Funding

This work is supported by the Natural Science Foundation of Chongqing (CSTB2022NSCQ-MSX0398, CSTB2022NSCQ-MSX1425), Science and Technology Foundation of the Education Department of Chongqing (KJQN202101510).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Zhao C, Sun Q, Zhang C, Tang Y, Qian F. Monocular depth estimation based on deep learning: An overview. *Sci China Technol Sci* (2020) 63(9):1612–27. doi:10.1007/s11431-020-1582-8

2. Ming Y, Meng X, Fan C, Yu H. Deep learning for monocular depth estimation. *A Review Neurocomputing* (2021) 438:14–33.

3. Liu X, Xue N, Wu T. Learning auxiliary monocular contexts helps monocular 3D object detection. *Proc AAAI Conf Artif Intelligence* (2022) 36:1810–8. doi:10.1609/aaai.v36i2.20074

4. Luo S, Dai H, Shao L, Ding Y. M3dssd: Monocular 3d single stage object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (2021). p. 6145–54.

5. Bhattacharyya S, Shen J, Welch S, Chen C. Efficient unsupervised monocular depth estimation using attention guided generative adversarial network. *J Real-Time Image Process* (2021) 18(4):1357–68. doi:10.1007/s11554-021-01092-0

6. Ye X, Fan X, Zhang M, Xu R, Zhong W. Unsupervised monocular depth estimation via recursive stereo distillation. *IEEE Trans Image Process* (2021) 30:4492–504. doi:10.1109/tip.2021.3072215

7. Sun Q, Tang Y, Zhang C, Zhao C, Qian F, Kurths J. Unsupervised estimation of monocular depth and VO in dynamic environments via hybrid masks. *IEEE Trans Neural Networks Learn Syst* (2021) 33(5):2023–33. doi:10.1109/tnnls.2021.3100895

8. Garg R, Bg VK, Carneiro G, Reid I. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: *European conference on computer vision*. Cham: Springer (2016). p. 740–56.

9. Godard C, Mac Aodha O, Brostow GJ. Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE (2017). p. 270–9.

10. Zhou T, Brown M, Snavely N, Lowe DG. Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE (2017). p. 1851–8.

11. Ding X, Guo Y, Ding G, Han J. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV); October 2019. IEEE (2019). p. 1911–20.

12. Liu S, Huang D. Receptive field block net for accurate and fast object detection. In: Proceedings of the European conference on computer vision (ECCV). IEEE (2018). p. 385–400.

13. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE (2018). p. 7132–41.

14. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); December 2016. IEEE (2016). p. 2921–9.

15. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR). IEEE (2017). p. 2881–90.

16. Liu C, Zhu L, Belkin M. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Appl Comput Harmonic Anal* (2022) 59:85–116. doi:10.1016/j.acha.2021.12.009

17. Mur-Artal R, Montiel JMM, Tardos JD. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans robotics* (2015) 31(5):1147–63. doi:10.1109/tro.2015.2463671

18. Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. *Adv Neural Inf Process Syst* (2014) 27:2366–74.

19. Liu M, Salzmann M, He X. Discrete-continuous depth estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 2014; Columbus, OH, USA. IEEE (2014). p. 716–23.

20. Li R, Wang S, Long Z, Gu D. Undeepvo: Monocular visual odometry through unsupervised deep learning. In: Proceedings of the 2018 IEEE international conference on robotics and automation (ICRA); May 2018; Brisbane, QLD, Australia. IEEE (2018). p. 7286–91.

21. Yang Z, Wang P, Xu W, Zhao L, Nevatia R. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. *Proc AAAI Conf Artif Intelligence* (2018) 32:12257. doi:10.1609/aaai.v32i1.12257

22. Dovesi PL, Poggi M, Andraghetti L, Martí M, Kjellström H, Pieropan A, Mattoccia S. Real-time semantic stereo matching. In: Proceedings of the 2020 IEEE international conference on robotics and automation (ICRA); Paris, FranceMay 2020. IEEE (2020). p. 10780–7.

23. Kundu JN, Uppala PK, Pahuja A, Babu RV. Adadepth: Unsupervised content congruent adaptation for depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; March 2018. IEEE (2018). p. 2656–65.

24. Chen X, Wang Y, Chen X, Zeng W. S2r-depthnet: Learning a generalizable depth-specific structural representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 2021; Nashville, TN, USA. IEEE (2021). p. 3034–43.

25. Yin Z, Shi J. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition; June 2018. IEEE (2018). p. 1983–92.

26. 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 (CVPR); June 2018. IEEE (2018). p. 5667–75.

27. Yang Z, Wang P, Wang Y, Xu W, Nevatia R. Lego: Learning edge with geometry all at once by watching videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition; June 2018. IEEE (2018). p. 225–34.