Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 3, pp. 1529–1538, DOI 10.12694/scpe.v26i3.3919

DEPTH ESTIMATION OF MONOCULAR VR SCENES BASED ON IMPROVED ATTENTION COMBINED WITH DEEP NEURAL NETWORK MODELS

GUANG HU *AND PEIFENG SUN†

Abstract. The boundary blurring issue with the existing unsupervised monocular depth estimation techniques is addressed by a suggested network design based on a dual attention module. This architecture is able to overcome the boundary blurring issue in depth estimation by making effective use of the remote contextual information of picture features. The model framework comprises of a pose estimation network and a depth estimation network to estimate depth and camera pose transformations simultaneously. The complete framework is trained using an unsupervised method based on view synthesis. The depth estimation network incorporates a dual attention module, comprising a position attention module and a channel attention module. This allows the network to estimate the depth information more precisely by representing the distant spatial locations and the contextual information between various feature maps. Based on the KITTI and Make3D datasets, the experimental findings demonstrate that this method may successfully solve the depth estimation border ambiguity problem and increase the accuracy of monocular depth estimation.

Key words: Self-Attention Mechanism, Monocular Depth Estimation, Photometric Loss, Image Reconstruction,Depth Estimation Accuracy.

1. Introduction. Depth information plays an important role in understanding 3D scenes and it can be applied to various robotics techniques such as 3D reconstruction, 3D target detection and Simultaneous Localization and Mapping (SLAM) [1]. The task of obtaining depth information from an image is known as image depth estimation, and recovering pixel-level depth through images is gaining interest in the field of computer vision due to properties such as lightness and cheapness of cameras [2, 3].

With the rapid development of deep learning techniques, many works use supervised depth learning to infer depth information from images. However, the acquisition of truth data required for supervised learning is not easy, so recent work attempts to solve the depth estimation problem using unsupervised learning [4]. To learn the mapping from pixels to depth in the absence of true annotations, the model needs to have other constraints attached. One form of unsupervised depth estimation is to use synchronized binocular image pairs for training [5]. The simultaneous binocular image pairs are used only during training, and the model estimates the left-right image parallax or image depth, thereby reconstructing the image by comparing the image The model is trained by comparing the differences between the images [6].

For the study of monocular image depth estimation, a large number of research methods have been proposed by domestic and foreign researchers in this direction [7]. In recent years, the rise of deep learning has also had a great impact on the field of deep estimation, and many research methods based on deep learning have been proposed with excellent results. Three popular types of methods for image depth estimation are currently available-supervised learning methods, joint semantic segmentation methods, and unsupervised learning [8]. Models are trained using supervised learning, and the training uses datasets labeled with a large amount of depth information. Two networks are overlaid: the first network is the Global Coarse-Scale Network, which performs coarse-scale global prediction of images; the other network is the Local Fine-Scale Network, which is mainly responsible for local refinement. The performance is improved by CRF normalization. The basic idea of this study is to use multi-scale neural networks to estimate the depth map [9]. It proposed a model with discrete depths for the problem of new view synthesis and subsequently extended this approach by estimating continuous parallax values.[10] produces better results than current partially supervised methods by using a left-right depth consistency term. Another unsupervised form with fewer constraints is to use monocular video

*Computer Department, Zhengzhou Preschool Education College, Zhengzhou, Henan 450099, China. (huguang616@126.com).
†Computer Department, Zhengzhou Preschool Education College, Zhengzhou, Henan 450099, China.

data to train the model, using image reconstruction losses as a supervised signal to train the network. This unsupervised training approach requires the network to estimate the camera pose between frames in addition to the estimated depth. [11] pioneered the use of only monocular video to train a depth estimation network as well as a separate bit-pose estimation network. To handle non-rigid scene motion, they proposed to use the network to learn to interpret the mask, allowing the model to ignore specific regions that violate the rigid scene assumption. [12] used a more explicit geometric loss to jointly learn depth and camera motion for rigid scenes. A refined network was added to the study of in the literature to estimate the residual optical flow. These methods accomplish the training task using only monocular video sequences or binocular image pairs and produce better results than partially supervised methods in outdoor scenes [13].

However, none of the above methods make good use of the contextual information in the scene. [14]studied the statistics of depth images of natural scenes and showed that depth images can be decomposed into segmented smooth regions with little dependence on each other and often with sharp discontinuities. Therefore, the variation of scene depth is closely related to the concept of ”object” in the scene, rather than some underlying features like color, texture, illumination, etc. Some of the current studies [15] use edge-aware smoothing loss to constrain the model to produce a smoother depth image within the ”object”. However, the edge map based on image gradient does not represent the object boundary well. To solve this problem, this paper proposes to improve the depth estimation network using the dual attention module proposed in [16] in the field of semantic segmentation to enhance the feature extraction capability of the model by using the intra- and inter-object contextual information more effectively through the attention mechanism. The validation results of this paper's approach on the KITTI dataset and Make3D dataset demonstrate the effectiveness of the attention mechanism in improving the depth estimation accuracy.

Here are the major contributions of our paper:

This paper introduces a dual attention module combining spatial and channel attention mechanisms, significantly enhancing the model's ability to capture both local and global context in monocular unsupervised depth estimation.

Through the integration of self-attention mechanisms, the proposed model demonstrates superior performance in terms of error reduction and threshold accuracy on the KITTI dataset, outperforming several state-of-the-art methods.

A robust photometric loss function combining Structural Similarity Index (SSIM) and L1 parametrization is designed to address illumination effects and enhance view reconstruction accuracy.

2. Literature Review.

Monocular depth estimation has gained significant attention in recent years due to its wide range of applications in autonomous driving, augmented reality, and scene understanding. This section reviews several recent studies in the field, highlighting their contributions and how the proposed work in this paper compares to them.

2.1. Traditional Depth Estimation Approaches.

Early works on depth estimation primarily relied on supervised learning techniques, requiring large datasets with ground truth depth information. For instance, [7] developed one of the earliest multi-scale convolutional neural network (CNN) models for depth estimation, using a coarse-to-fine approach to predict depth at various scales. However, supervised methods face challenges due to the scarcity of labeled data and their reliance on high-cost depth sensors for ground truth data collection.

2.2. Unsupervised Learning Methods.

To overcome the limitations of supervised approaches, unsupervised methods have been proposed that rely on stereo image pairs or monocular sequences for training without ground truth labels. [5] introduced an unsupervised method using stereo images, leveraging a photometric loss based on image reconstruction. Their method greatly reduced the need for expensive depth sensors but suffered from limitations related to image occlusions and moving objects.

The paper [10] further advanced this area by introducing a fully unsupervised framework using only monocular video sequences. They introduced a view synthesis approach that allowed the network to learn depth estimation without stereo pairs, making the approach more generalizable. Despite these advancements, their method struggled with capturing fine details and often produced artifacts in object boundaries.

2.3. Attention Mechanisms in Depth Estimation.

Recently, attention mechanisms have been integrated into depth estimation models to enhance feature extraction and focus on important regions of the image.

[13] incorporated a spatial attention module to improve scene understanding, demonstrating improved accuracy on the KITTI dataset. However, their approach lacked an effective strategy to capture channel dependencies, which limited the model's ability to fully leverage multi-channel feature maps.

Incorporating both spatial and channel attention, [15] proposed an approach to improve the accuracy of depth estimation by enhancing the model's ability to capture the relationships between different feature channels. While their approach demonstrated superior performance, the model still faced difficulties in preserving fine-grained details, particularly in complex scenes with occlusions.

2.4. Recent Developments. Recent works such as those by [4] and [6] have further advanced the field by introducing novel architectures and loss functions to improve depth estimation accuracy. [11] presented a multi-scale feature fusion approach that improved the network's ability to generalize across different datasets, while [12] explored depth estimation in diverse scenarios using large-scale datasets. Both approaches improved generalization but did not address the issue of enhancing feature compactness within objects and improving overall feature distinguishability.

Compared to recent works, our approach demonstrates a more balanced and robust framework for monocular depth estimation, addressing limitations in both contextual understanding and feature preservation. By leveraging dual attention mechanisms and a robust loss function, the proposed method outperforms state-of-the-art models in terms of both error reduction and depth prediction accuracy, particularly on challenging datasets like KITTI.

3. Related Research.

3.1. Problem description. The task of predicting the scene depth from the image data is known as depth estimation of the image [17]. The image captures the projection information of the three-dimensional world on the imaging plane, It falls under the category of computer-related 3D reconstruction, and this issue is expressed mathematically as $D = F(I)$, where is D depth, I is the image, and F is the mapping function from the image to the depth. Monocular depth estimate is an ill-posed (ill-posed) problem because of the ambiguity of the scale, so it can hardly be solved directly F . Many scholars have started to use supervised deep learning for depth estimation, however, because gathering large-scale, real-labeled data is costly and time-consuming, a lot of recent research has concentrated on unsupervised deep learning techniques.

3.2. View reconstruction as a supervised signal. Using view reconstruction as a supervised signal is an unsupervised method, and its core idea is to use depth and pose as intermediate quantities, combined with pairwise polar geometry for view reconstruction. Assuming that the observation scene is stationary, given two views taken at different viewpoints I_t, I_s , if the coordinate transformation matrix of the depth map D_t, I_t to the view I_t is known, the pixel mapping relationship between I_t, I_s .

$$p_s = KT_{t\sim s}D_tK^{-1}p_t \tag{3.1}$$

where K is the camera internal reference, $T_{t\sim s}$ is the coordinate transformation matrix from I_t to I_s, and p_t, p_s are the pixel coordinates of the two views, respectively. The network model can learn the interframe posture transformation and the depth of each pixel, so that the images from different views can be synthesized and compared with the target view using an interpolation algorithm (e.g., bilinear interpolation) based on the mapping relationship in Eq.3.1, and thus the depth and pose transformation can be estimated by unsupervised training of the model.

4. System Model Framework.

4.1. Network structure overview. As can be observed in Fig.4.1, the bit-pose transformation estimate network and the depth estimation network are the two parts of the model framework used in this work. In this paper, a single color image is used as the input for the depth estimation network. Its result is a dense depth map, which is different from some earlier research. Moreover, the training of the entire system is easier to converge since direct depth estimation involves less inverse operations than parallax estimation. Two pictures are fed into the bit-pose estimation network, and a 6-Do F bit-pose transform is produced as the output. The training process does not require the real depth and the pose-transform annotation of the actual camera motion.

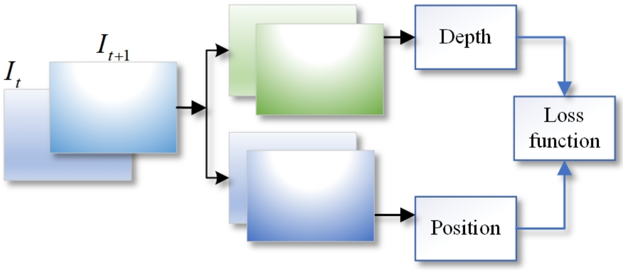

Fig. 4.1: Model framework.

Instead, the depth map estimated by the model and the pose-transform are used for view reconstruction, and the contrast error between the reconstructed view and the target view is used as a loss to train the neural network.

Two fully convolutional networks—a depth estimation network and a bit-pose transform estimation network—make up the model architecture in this work. The structure of the depth estimation network, which is based on the U-Net architecture, is depicted in Fig.4.2. In order to represent distant contextual information while extracting deep features, it incorporates jump connections and attention modules. To extract strong image characteristics, this paper uses ResNet18 as the encoder for RGB picture feature extraction. Compared to the encoders in earlier studies that employed Disp Net and Res Net50-based models, the encoder in this study operates more quickly and requires less parameters. In this study, pre-trained weights from Image Net are used to initialize the encoder weights. Tests show that as compared to training the model from scratch, this initialization improves accuracy.

Since the encoder downsamples the input image to extract the feature map, an upsampling procedure is required to perform the feature map resolution reduction. The decoder of the deep estimation network, consisting of five upsampling modules, uses the Exponential Linear Unit (ELU) as the activation function everywhere except at the output. A convolutional operator layer and the nearest neighbor interpolation method make up the upsampling module of this paper. Fig.4.2 dashed-labeled area illustrates the construction of this module. In this paper, the attention module is added to the decoder section of the deep estimation network in order to model the remote contextual information and improve the correlation between features. To learn the contextual information between features without introducing too much computational overhead, a two-channel attention module—which consists of a location attention module and a channel attention module—is inserted in the first two layers of the decoder. The image's depth information is output via a Sigmoid activation function and a 3x3 convolution process, which make up the depth estimation layer. This study performs a linear transformation of the output to constrain it to a tolerable range.

The encoder portion of the bit-pose transform estimation network is a conventional Res Net18 structure, and the entire convolutional network with six input and output channels is used. The decoder consists of four layers of convolutional operations: layers 1 and 4 have 1×1 convolutional kernel sizes, while layers 2 and 3 have 3×3 convolutional kernel sizes. Rectified Linear Unit (ReLU) activation functions are present in all layers except the output layer. Image sequences are fed into the network via batch size stacking. The encoder then extracts the feature maps, and further convolution operations are used to derive the higher-level features of the various frames, and finally the output pose is output by 1×1 size convolution. The output bit-pose is a 6-dimensional bit-pose transformation vector, with the first 3 dimensions representing rotation and the last 3 dimensions representing displacement.

4.2. Depth estimation network combining dual attention module. Sometimes, the convolution technique breaks the depth estimation for some elongated objects (like streetlights) since it has a limited perception range and the object objects in the input image fluctuate in scale, angle, and brightness. In order to maximize the accuracy of the depth estimate and make better use of the global knowledge of the scene and the relationship between the representation properties, this research employs the dual attention module in the

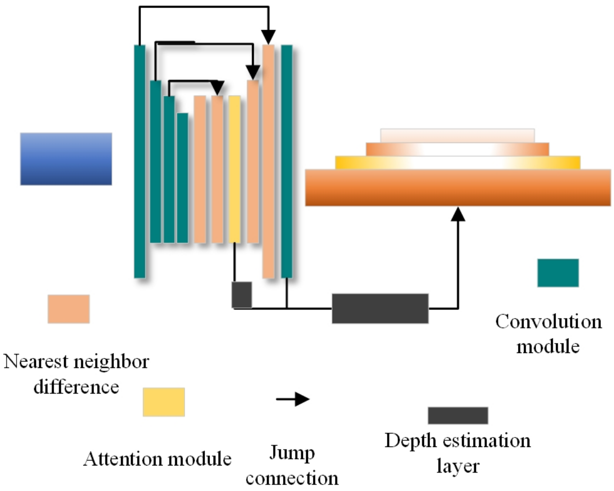

Fig. 4.2: Depth estimation network structure.

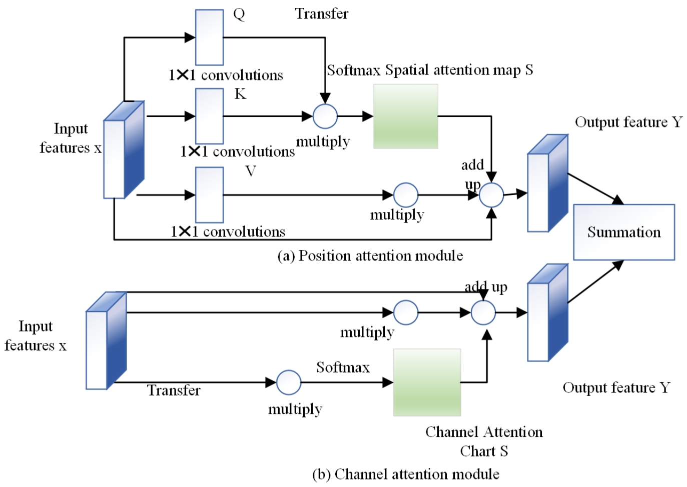

Fig. 4.3: Dual-focus module.

depth estimation network.

The location attention module and the channel attention module are the two attention modules that make up the dual-attention module. The spatial and channel characteristics of remote contextual information are captured by the two attention modules. The depth estimation network's decoder incorporates the dual-channel attention module, and Fig.4.3 displays a schematic of the two attention modules' structural layout.

4.2.1. Location attention module. Traditional complete convolutional networks are prone to the issue where the edges do not match the actual objects when estimating depth because they extract local features that lack global information to indicate the link between local features. This study presents the location attention module to model the contextual relationships of local features. For the feature map $X \in R^{C\times H\times W}$ encoded by the convolution layer, it is first fed into the 1×1 convolution layer to downscale the number of channels and generate two new features $Q \in R^{\frac{C}{r}\times H\times W}, K \in R^{\frac{C}{r}\times H\times W}$ respectively, where takes the value of 8 in this paper. Q, K are then reshaped into $Q \in R^{\frac{C}{r}\times N}, K \in R^{\frac{C}{r}\times N}$ and the transpose of Q is matrix multiplied with K , where $N = H \times W$. Finally, the obtained results are passed through the softmax layer to calculate the spatial attention map $S \in R^{N \times N}$, as shown in Eq.4.1 is shown.

$$S_{ji} = \frac{\exp(Q_i \cdot K_j)}{\sum_{i=1}^{N} \exp(Q_i \cdot K_j)} \tag{4.1}$$

The stronger the correlation between two locations, the more similar the feature representations of those sites are. In the meantime, a new feature map is created by feeding the input features X into the convolution layer . The V is reshaped into $V \in R^{C \times N}$ and then matrix multiplication is performed between the transpose of V and$Y_i = \alpha \sum\limits_{i=1}^{N} (S_{ji} V_i) + \beta X_{_j} \in X R^{C \times H \times W}, S_{ji} = \frac{\exp(Q_i \cdot K_j)}{\sum_{i=1}^{N} \exp(Q_i \cdot K_j)} \in R^{N \times N}, V \in R^{C \times N}, r, Q, K, N = H \times W$. Finally, to make the module more flexible, the result of multiplying V and S with the input features X is multiplied by the element-by-element summing operation and the scale parameter. is performed in this paper to obtain the final output $Y \in R^{C \times H \times W}$, as shown in Eq.4.2

$$Y_i = \alpha \sum_{i=1}^{N} (S_{ji} V_i) + \beta X_{_j} \tag{4.2}$$

where α is initialized to 0, β is initialized to 1, as the training eventually assigns both weights. From Eq.4.2, it can be derived that the output feature Y at each location is a weighted sum of the features at all locations and the original features. As a result, it collects contexts selectively using the spatial attention network and has a global context view. When similar features of an object are associated, the compactness of the features inside the object is enhanced.

4.2.2. Module for Channel Attention. The high-level feature map of each channel can be viewed as an object-specific response, and there are relationships between various feature maps that are intimately connected to the three-dimensional structure of the scene. A specific scene object's feature representation can be enhanced by the model by taking advantage of the interdependencies between channel feature mappings. Consequently, the channel attention module, the structure of which is depicted in Fig.4.3b, is used in this research to explicitly represent the interdependencies between channels. Here, the channel attention map is computed directly from the original characteristics, in contrast to the location attention module. Specifically, the input features $X \in R^{C \times H \times W}$ are reshaped into $X \in R^{C \times N}$ matrix multiplication between their transpose, and then the softmax layer is applied to obtain the channel attention map $S \in R^{C \times C}$, see Eq.4.3

$$S_{ji} = \frac{\exp(X_i \cdot X_j)}{\sum_{i=1}^{N} \exp(X_i \cdot X_j)} \tag{4.3}$$

where S_{ji} measures the effect of the i -th channel on the j -th channel. Subsequently, a matrix multiplication operation is performed between the transpose of S and X . The result is then multiplied with the input feature X by the scale parameter and subjected to an element-by-element summation operation to obtain the final output $Y \in R^{C \times H \times W}$, as shown in Eq.4.4:

$$Y_j = \lambda \sum_{i=1}^{C} (S_{ji} X_i) + \omega X_{_j} \tag{4.4}$$

where λ, ω learn the weights gradually starting from 0 and 1, respectively. After processing by the channel attention module, each channel's final feature is the weighted sum of its initial characteristics as well as the features of all other channels, It enhances feature distinguishability and aids in the network's representation of the scene's structural information by modelling the remote dependencies between feature mappings.

4.3. Loss function design. In this article, the model is trained using the difference between the synthetic image and the target view as a supervised signal, so the design of the image comparison loss function is an important part. Since the camera motion is easily affected by illumination, this paper uses the robust similarity

Table 5.1: Error results compared before and after the self-attention module was included.

Method	AbsRel	SqRel	RMSE	LogRMSE
This algorithm (without self attention mechanism)	0.097	0.796	4.631	0.199
This algorithm	0.091	0.717	4.415	0.181

Table 5.2: Comparison of threshold accuracy results before and after adding the self-attention module.

Method	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
This algorithm (without self attention mechanism)	0.858	0.944	0.978
This algorithm	0.881	0.959	0.981

comparison function in the literature [17] as the loss function of the model to judge the good or bad view reconstruction, i.e., the combination of Structural Similarity Index (SSIM) and L1 parametrization, and the specific photometric loss function:

$$L_p = \alpha \frac{1 - ssim(I_t I_t)}{2} + (1 - \alpha)|I_t - I_t| \tag{4.5}$$

where I is the real view, I_t is the synthetic view, and α is the weight parameter, which is set here to 0.85. In the image sequence, the image contrast luminosity loss can be obtained according to the loss function by using the images of moment $t-1$ and moment $t+1$, respectively, to synthesize the image of moment . In order to reduce the effect of occlusion and moving objects, this paper uses the minimum value of the synthetic loss of taking different frames as the final loss in the literature [17], that is

$$L = \frac{1}{N} \sum_{i=0}^{N} \min_{t} L_p \left(I_t I_{i \to t} \right) \tag{4.6}$$

Here L_p denotes the photometric loss function of Eq.4.5, and is the total number of pixels. Due to the bilinear interpolation with subdifferentiation, the loss is calculated for the output of the four scales in this paper so as to reduce its effect.

5. Analysis and outcomes of the experiment.

5.1. Quantitative analysis. This chapter deals with monocular picture depth estimation using unsupervised learning techniques. Comparative studies are carried out to confirm the algorithm's efficacy following the addition of the self-attention module to Attention-Unet. The experimental findings before and after the self-attention mechanism was added to the Attention-Unet network in the depth estimation network are compared in Table5.1 and Table5.2. The results of the experimental comparison data demonstrate that the estimation network functions better on the KITTI dataset when the attention mechanism is added.

The depth estimation network, which is composed of several attention modules, may now incorporate self-attention to better gather context about the image and avoid the problem of losing image object features in the network model during depth estimation. The network uses a large number of Skip-Connections at the same time, which can fuse all feature information and hasten the convergence of the network. This also enhances some invalid and sparse feature information, improving the performance of the network model. Following data comparison, the self-attention mechanism in the Attention-Unet network in the depth estimation network improves the model's error and threshold accuracy, and the result on the threshold accuracy of $\delta < 1.25$ is improved by 2.3% compared with the algorithm without the self-attention mechanism. The experiments will be compared with a few popular algorithms to confirm the efficacy of this approach. Table5.3 presents the comparison between the method used in this chapter and other methods that were trained on the KITTI dataset and subsequently tested on the Eigen Split test set.

Table 5.3: Error results compared with other methods.

Method	AbsRel	SqRel	RMSE	LogRMSE
Song [5]	0.218	1.777	6.857	0.279
Zhang [7]	0.199	1.549	6.301	0.278
Osamah [11]	0.176	1.171	5.286	0.278
An [13]	0.139	1.340	5.850	0.237
Li [15]	0.120	0.840	4.497	0.195
ZKaushik [17]	0.099	0.765	4.486	0.189
Our	0.089	0.728	4.411	0.183

Table 5.3 compares the experimental findings with the state-of-the-art methods; the suggested algorithm in this research yields the best results. A full-resolution image is used as the training input for the depth estimation network model, which is based on a dual network structure. For the extraction of global features the network with relatively deep depth is used to process the high-resolution scene images, and the relatively shallow network is used to process the low-resolution scene images to extract local detailed features. However, this method has the potential to lead to region estimation errors and local details missing in the image. Compared with this method, the results of the model in this paper have lower errors, with 4.1% improvement in the threshold accuracy of $\delta < 1.25$ and 0.6% improvement in the threshold accuracy of $\delta < 1.25^2$.

By using a self-attention module, the network model with joint attention mechanism presented in this paper is able to gather contextual information and detail information of the scene images more effectively. In the comparison of the threshold accuracy results for $\delta < 1.25$, the results of the method in this paper improve 0.6% and 0.3% in all the accuracies of $\delta < 1.25^2$,$\delta < 1.25^3$. As a result, the technique presented in this paper improves in error as well as thresholding accuracy compared to other popular algorithms.

5.2. Qualitative Analysis. Experiments add other modules to the network independently and perform control experiments on the same dataset in order to further validate the methodology presented in this paper. The results are displayed in Fig.5.1. Three scenes are chosen for comparison experiments: a) the row depicts the scene as it was originally seen; b) the row employs the most basic network structure without including the attention mechanism, automatic masking loss function, and combined reprojection loss; and c) the depth map derived from the experiment is distorted by numerous artifacts. c) segments the image and drastically reduces the artifacts in the graph by using the suggested network with reprojection loss and automatic masking loss function for training without the attention method. However, there are still some errors, and the tree trunk in scene A and the outline of the column in scene B with the trees in the distance and the trees on the left in scene C are not yet completely clear. d) Row method, i.e., the network structure proposed in this paper, adds attention mechanism to the depth estimation network, and after combining reprojection loss and automatic masking, the effect of depth map is further improved, e.g., in scene B, the excess shadow contour on the top of the column is removed, and a more accurate presentation of the column contour is obtained, while the obscured trees, vehicles, etc. are presented with a clearer effect.

The experimental results demonstrate that the automatic masking loss function and reprojection loss may successfully decrease the artifacts caused by moving objects. Additionally, the depth map formed with the self-attention mechanism has a higher hierarchical structure and is more clearly delineated. It has been shown that integrating the attention mechanism with the automatic masking block, reprojection loss, and other components improves the performance of the depth estimation network.

Two scenarios are re-selected in order to compare the outcomes before and after utilising the self-attention mechanism (self-Attention) in the depth estimation network architecture with other conditions remaining consistent. The comparison plots are displayed in Fig.5.2, where it is evident that the depth prediction is improved over that which would have resulted from the absence of the self-attention mechanism. After the self-attention mechanism is added, the contours of automobiles and trees are more easily distinguished, shadows are lessened, and the outlines of objects that are relatively far away are more clearly defined and clearly layered.

The depth estimation results are compared with those of the algorithm in the literature [5] on the KITTI

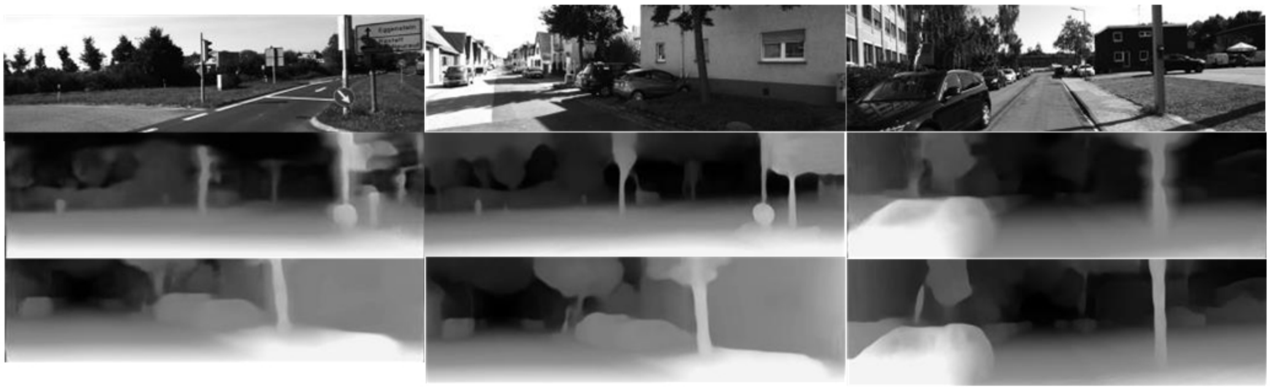

Fig. 5.1: Comparison of the results of adding different modules.

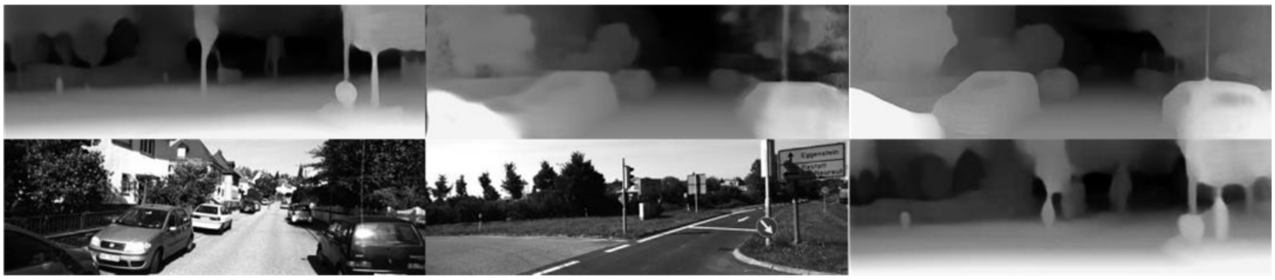

Fig. 5.2: Comparison of before and after adding attention mechanism.

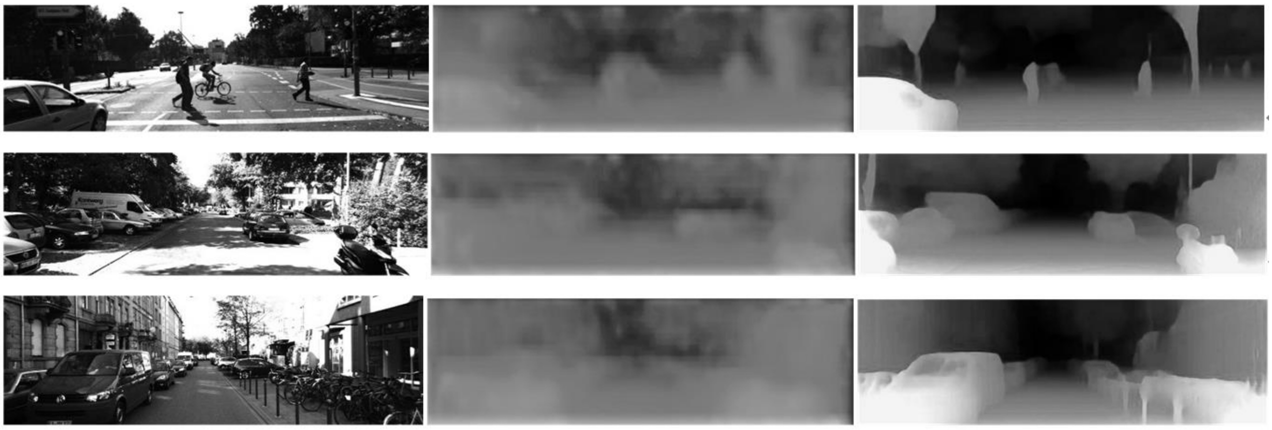

Fig. 5.3: Comparison of the depth estimation results with the literature [5].

Eigen Split test set in order to confirm the efficacy of the approach in this paper. The comparison graph is displayed in Fig.5.3. The scene maps in the figure are obtained from the KITTI dataset, and this experiment is conducted to compare three scenes separately, and it is evident from the three scene maps that this paper's depth estimation map performs better than the literature's method in terms of overall depth estimation, tiny item recognition, and hierarchical separation contouring.

6. Conclusion. This paper presents a novel unsupervised monocular depth estimation method based on a dual attention mechanism. The incorporation of both spatial and channel attention modules allows the network to effectively capture global contextual information and enhance the structural details of the depth map. Experimental results on the KITTI and Make3D datasets demonstrate that the proposed method achieves superior accuracy compared to existing approaches. By addressing the challenges of object feature loss and improving depth prediction for complex scenes, the model exhibits strong generalization capability. Future work will focus on optimizing the pose estimation network and integrating binocular cues to further enhance depth estimation accuracy.

Data Availability. The experimental data used to support the findings of this study are available from the corresponding author upon request.

REFERENCES

[1] Zhu, S. ,& Zhao, H. *Depth estimation of monocular infrared images based on attention mechanism and graph convolutional neural network.* Journal of Applied Optics, 42(1),(2021) 49-56.

[2] Chen, Y. , Zhao, H. , Hu, Z. , & Peng, J. *Attention-based context aggregation network for monocular depth estimation.* International Journal of Machine Learning and Cybernetics(11),(2021)1-14.

[3] Lei, Z., Wang, Y., Li, Z., & Yang, J. *Attention based multilayer feature fusion convolutional neural network for unsupervised monocular depth estimation.* Neurocomputing, 423,(2021) 343-352.

[4] Liu, P., Zhang, Z., Meng, Z., & Gao, N. *Joint attention mechanisms for monocular depth estimation with multi-scale convolutions and adaptive weight adjustment.* IEEE Access, 8,(2020) 184437-184450.

[5] Song, M., Lim, S., & Kim, W. *Monocular depth estimation using laplacian pyramid-based depth residuals.* IEEE transactions on circuits and systems for video technology, 31(11),(2021) 4381-4393.

[6] Song, X., Li, W., Zhou, D., Dai, Y., Fang, J., Li, H., & Zhang, L. *MLDA-Net: multi-level dual attention-based network for self-supervised monocular depth estimation.* IEEE Transactions on Image Processing, 30,(2021) 4691-4705.

[7] Zhengwan, Z. H. A. N. G., Chunjiong, Z. H. A. N. G., Hongbing, L. I., & Tao, X. I. E. *Multipath transmission selection algorithm based on immune connectivity model.* Journal of Computer Applications, 40(12),(2020) 3571. DOI: 10.11772/j.issn.1001-9081.202004049.

[8] Cheng, Z., Zhang, Y., & Tang, C.*Swin-Depth: Using Transformers and Multi-Scale Fusion for Monocular-Based Depth Estimation.* IEEE Sensors Journal, 21(23),(2021) 26912-26920.

[9] Xiang, X., Kong, X., Qiu, Y., Zhang, K., & Lv, N. *Self-supervised Monocular Trained Depth Estimation Using Triplet Attention and Funnel Activation.* Neural Processing Letters, 53(6),(2021) 4489-4506.

[10] He, L., Lu, J., Wang, G., Song, S., & Zhou, J.*SOSD-Net: Joint semantic object segmentation and depth estimation from monocular images.* Neurocomputing, 440,(2021) 251-263.

[11] Osamah Ibrahim Khalaf, Carlos Andrés Tavera Romero, Shahzad Hassan, Muhammad Taimoor Iqbal, *"Mitigating Hotspot Issues in Heterogeneous Wireless Sensor Networks", Journal of Sensors, vol.* 2022, Article ID 7909472, 14 pages, 2022. https://doi.org/10.1155/2022/7909472.

[12] Khaparde, A. R., Alassery, F., Kumar, A., Alotaibi, Y., Khalaf, O. I. et al. *Differential Evolution Algorithm with Hierarchical Fair Competition Model.* Intelligent Automation & Soft Computing, 33(2), 1045–1062. doi:10.32604/iasc.2022.023270.

[13] An, P., Wang, Z., & Zhang, C. *Ensemble unsupervised autoencoders and Gaussian mixture model for cyberattack detection.* Information Processing & Management, 59(2),(2022) 102844.

[14] Naveed Ahmad Khan, Osamah Ibrahim Khalaf, Carlos Andrés Tavera Romero, Muhammad Sulaiman, Maharani A. Bakar, *"Application of Intelligent Paradigm through Neural Networks for Numerical Solution of Multiorder Fractional Differential Equations", Computational Intelligence and Neuroscience, vol.* 2022, Article ID 2710576, 16 pages, 2022. https://doi.org/10.1155/2022/2710576.

[15] Li, Y., Luo, F., Li, W., Zheng, S., Wu, H. H., & Xiao, C. *Self-supervised monocular depth estimation based on image texture detail enhancement.* The Visual Computer, 37(9),(2021) 2567-2580.

[16] Bhattacharyya, S., Shen, J., Welch, S., & Chen, C. *Efficient unsupervised monocular depth estimation using attention guided generative adversarial network.* Journal of Real-Time Image Processing, 18(4),(2021) 1357-1368.

[17] ZKaushik, V., Jindgar, K., & Lall, B. *ADAADepth: Adapting Data Augmentation and Attention for Self-Supervised Monocular Depth Estimation.* IEEE Robotics and Automation Letters, 6(4),(2021) 7791-7798.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: May 13, 2024
Accepted: Feb 11, 2025