2022 2nd International Conference on Algorithms, High Performance Computing and Artificial Intelligence (AHPCAI) | 979-8-3503-3478-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/AHPCAI57455.2022.10087639

Monocular Depth Estimation with Multiscale Feature Fusion Networks

Yuhang Zheng[1], Chuqing Cao[1,2*]
[1]School of computer and information, Anhui Polytechnic University, Wuhu, Anhui, China
[2] Wuhu Institute of Robot Industry Technology, Harbin Institute of Technology, Wuhu, Anhui, China
Email: zhengyuhang0430@126.com

Abstract. Monocular image depth estimation has some problems, such as fuzzy depth estimation, inaccurate distance information and incomplete details in complex scenes. Aiming at these problems, a monocular depth estimation method based on pyramid vision transformer network optimization is proposed. An encoder with a pyramid transformer as the skeleton network is used to segment the image and obtain the position information between each pixel block, while a lightweight decoder is used and feature fusion is improved. Experiments on the dataset demonstrate that the proposed network can enhance the edge details and improve the accuracy of depth estimation.

Keys: monocular depth estimation; Adaptive Selection Feature Fusion; pyramid Vision transformer

I Introduction

Although the depth estimation technology using lidar and binocular cameras is relatively mature, it is difficult to work in special environments, and the problems of complex depth estimation technology and high hardware requirements still exist[1]. Monocular depth estimation requires little hardware and can adapt to complex application scenarios. For the high cost of using RGB-D cameras in visual SLAM, the effective ranging range is small, and the depth information error of using binocular cameras is large. Visual SLAM for depth estimation using a monocular camera due to the disadvantages of computational complexity. However, in the process of using a single image to predict the depth, the problem of depth information estimation error and insufficient precision will inevitably occur[2]. Therefore, the monocular depth estimation must solve the problem of low accuracy to deal with complex environments.

As convolutional neural networks are used for depth estimation, the effect is remarkable[3] . EIGEN [4] et al. used a multi-scale depth prediction technique that roughly estimates the depth globally and refines the estimated depth locally. Next, LONG [5] et al. proposed the use of a full-volume machine network (FCN) for semantic segmentation which is widely used in image prediction tasks, including image depth prediction. Subsequently, CHEN [6] et al. went a step further in the exploration of multi-scale information for monocular depth estimation. They consider the underlying scene structure information at different scales and use the RESNET residual network combined with SENET to adaptively select features from all scale features through an attention mechanism and use them for residual depth estimation at different structural scales. KIM [7] et al. used a PVT v2 transformer as the encoder skeleton network to obtain global and local features and used a lightweight decoder to predict depth. These methods all adopt the encoder-decoder method based on the fully connected layer, but in the process of depth estimation through local and global features, the feature information will inevitably be lost or the more important feature information will be ignored[8], resulting in less and less depth information acquisition Obscure outlines of objects, etc. Therefore, on the basis of the above research methods of monocular image depth estimation, this paper proposes a network structure of an encoder based on PVT v2 and a lightweight decoder with an adaptive feature fusion module. This method can improve the accuracy of depth estimation for the problems of unclear boundary contour and large errors of depth information. The sum of the scale-invariant logarithmic loss function and mean square loss function is used to reduce the error.

II Materials and Methods

Based on the idea of the Pyramid Vision Transformer by Wang et al. [9], this paper adopts Pyramid Vision Transformer (PVT v2) as the encoder of the backbone network. The transformer encoder in this paper enables the model to learn global features and the relationship between features, and the decoder can be optimized through residual connections and adaptive feature fusion modules' Local details. The network structure diagram is shown in Figure 1. This paper will introduce the proposed monocular depth estimation optimization algorithm from the design of the encoder, decoder and loss function.

A. Encoder

The network structure is shown in Figure 1. In the encoding stage, the feature map of the RGB image is first extracted by filtering. The subsequently extracted feature blocks are input to the spatial reduction attention (SRA) layer with attention mechanism and the MLP-DWconv-MLP layer with residual blocks. Then stitching is performed in an overlapping convolution manner to generate multi-scale features of different dimensions. This paper uses four transformer blocks, each of which generates scale feature maps of different dimensions.

B. Lightweight Decoder

This paper builds a lightweight decoder that uses fewer convolutions than traditional decoders using bilinear upsampling and stacked convolutional or deconvolutional layers and still achieves very good depth estimation results. The image feature size output by the encoder is H×W×C, and the dimension C is first reduced by 1×1 convolution. The feature map after continuous upsampling is fused with the feature map of the same resolution through the adaptive

979-8-3503-3478-4/22/$31.00 ©2022 IEEE

selection feature map module. After upsampling and repeating the above fusion and upsampling, the final output passes through two convolutional layers and a sigmoid function to generate the predicted Depth map H×W×1.

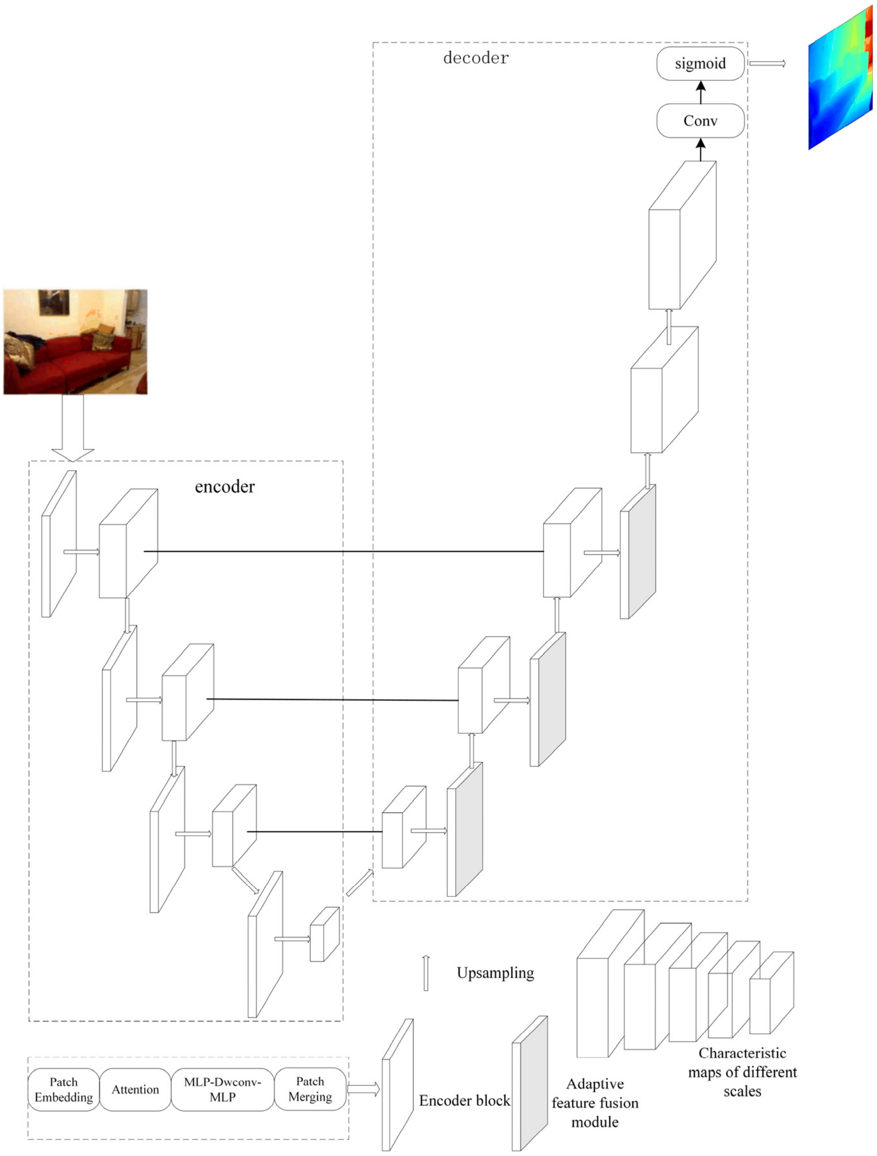

Figure 1. Pyramid Structure Optimization Neural Network Diagram

C. Adaptive Selection Feature Fusion

This paper proposes an Adaptive Selection Feature Fusion (ASFF) module, which uses adaptive learnable weight coefficients to assign learnable weights to the local and global feature maps during feature fusion [10], and then passes an attention map for each feature is obtained to adaptively select and integrate local and global features. The detailed structure of ASFF is shown in Figure 2. The feature fusion formula is as follows:

$$\mathrm{F} = \alpha_1 * Gconv + \alpha_2 * Lcon \qquad (1)$$

$$\mathrm{a_i} = \frac{\mathrm{e^{wi}}}{\sum \mathrm{je^{wj}}} (\mathrm{i} = 1{,}2; \mathrm{j} = 1{,}2) \qquad (2)$$

Among them, α_i is the normalized weight, $\sum\alpha_i = 1$, w is the initialization weight coefficient, and Gconv and Lconv represent the global feature map and local feature map. The final convolutional layer and sigmoid layer generate a two-channel attention map.

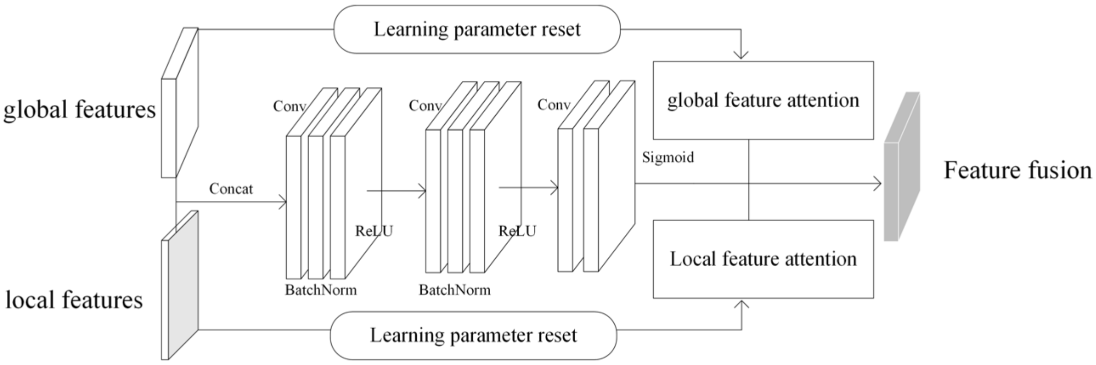

Figure 2. The detailed structure of adaptive feature fusion

D. Data Augmentation

In addition to the conventional center cropping, rotation and color gamut transformation, the data augmentation method in this paper also uses a depth-specific data augmentation method Cut Depth [11] to replace part of the RGB image with real depth information. Images provide variety and focus the network on high-frequency regions.

E. Loss calculation

A scale-invariant log-scale loss function is used to calculate the error between the predicted output and the true depth map[12]. The loss equation is:

$$\mathrm{Loss_{lg}} = \frac{1}{n}\sum_i d_i^2 - \frac{1}{2n^2}\left(\sum_i d_i^2\right) \tag{3}$$

$$d_i = log y_i - log y_i^* \tag{4}$$

where y_i, y_i^* represent the true value and the predicted value, respectively. In order to make the predicted value closer to the actual value, to minimize the difference, a mean square loss function is used[13]. The loss equation is:

$$\mathrm{Loss_{mse}}(y_i, y_i^*) = (y_i - y_i^*)^2 \tag{5}$$

The total loss function calculation equation is:

$$Loss = Loss_{lg} + Loss_{mse}$$

III Results and Discussion

To validate the optimization algorithm in this paper, our model is compared with existing methods through quantitative and qualitative evaluations. The designed deep neural network is built with PyTorch.

A. Test results and analysis on the NYU Depth v2 dataset

This paper uses the mainstream NYU Depth V2 dataset for training. In the experiment, 24,228 images are used for training and 654 images are used for evaluation. In this dataset, the training duration is 18h, and the training batch size is set to 12; the single-cycle learning strategy with Adam optimizer is used, the size of the initial learning rate is set to 3e-5, and the learning decay rate is 0.9, the learning rate is increased from 3e-5 to 1e-4 in the first half, the learning rate is reduced from 1e-4 to 3e-5 in the second half, and the Epoch is set to 25. The total training time of the algorithm in this paper on the NYU Depth v2 dataset is about 18h. The effect of the depth estimation method in this paper is evaluated by the following indicators: absolute relative error Abs Rel, root mean square error RMSE, error Log 10 and accuracy δ ($\delta<1.25$, $\delta<1.25^2$, $\delta<1.25^3$).

Table 1: Comparison of training network data on the NYU Depth V2 dataset

Method	AbsRel	RMSE	Log10	**δ**<1.25	**δ**<1.25^2	**δ**<1.25^3
Fu et al. [2018]	0.115	0.509	0.051	0.828	0.965	0.992
Eigen et al. [2014]	0.158	0.641	_	0.769	0.950	0.988
Chen et al. [2019]	0.111	0.514	0.048	0.878	0.977	0.994
Kim et al. [2022]	0.125	0.373	0.048	0.859	0.971	0.989
Wang et al. [2022]	0.119	0.411	--	0.886	0.975	0.995
Ours	0.111	0.372	0.046	0.861	0.974	0.993
Ours + MSE Loss	0.109	0.368	0.045	0.887	0.974	0.996

As can be seen from Table 1, the comparison of the experimental results before and after adding MSE loss shows that after adding MSE Loss to the depth estimation network, the performance on the NYU Depth dataset is better. The algorithm in this paper is superior to the existing depth estimation algorithms in terms of accuracy improvement and error reduction in monocular depth estimation.

The experimental visualization results are shown in Figure 3. Figure 4 is a comparison of the enlarged local details of the visualization results of different literature methods. Through the comparison of the results in Figure 3 and Figure 4, it can be found that the method in this paper is rich in local details and clearer. The displayed object contour can meet the use of visual SLAM.

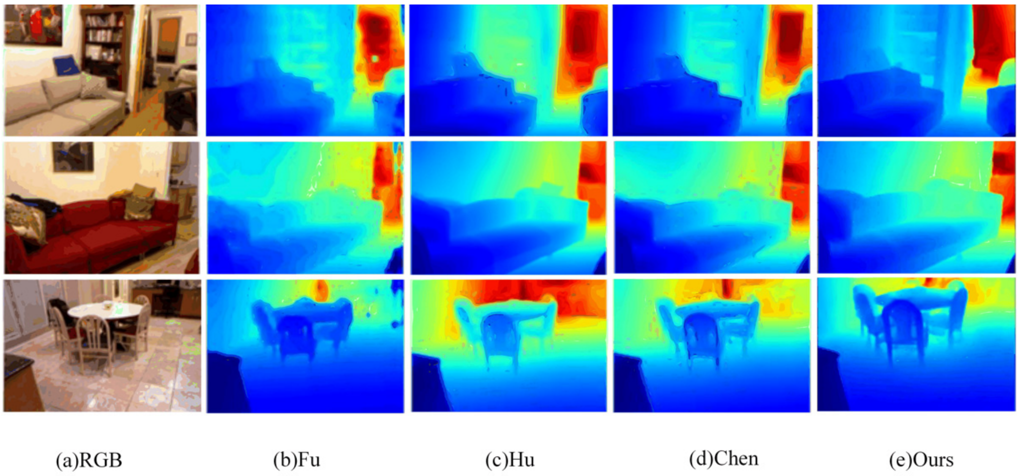

Figure 3. Visualization results of each network on the NYU Depth v2 dataset

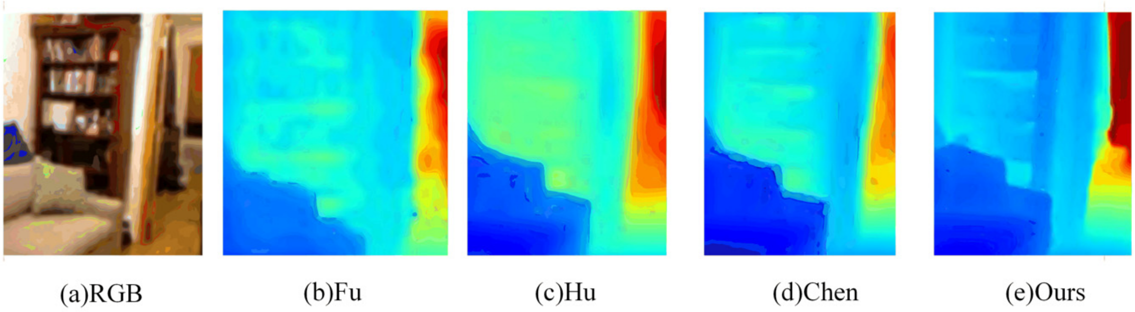

Figure 4. Comparison of local details

B. *Test results and analysis on the KITTI dataset*

Using the outdoor dataset KITTI dataset to train and evaluate the network can verify the generalization of the network in this paper. In the experiment, 23148 images are used for training the model, and 652 images are used for testing and evaluation. The experimental parameters are consistent with the above experiments. A maximum depth of 80 m was used for evaluation according to standard practice. The experimental evaluation data adds the square root relative error Sq Rel, logarithmic root mean square error log RMSE on the basis of the previous experiment Removed Log10.

Table 2. Comparison of evaluation results on the KITTI dataset

Method	AbsRel	SqRel	RMSE	LogRMSE	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Ruan et al. [2022]	0.147	1.26	5.74	0.242	0.789	0.924	0.967
Groenendijk et al. [2020]	0.152	1.36	6.00	0.249	0.788	0.917	0.963
Yang et al. [2020]	0.099	0.763	4.485	0.185	0.885	0.958	0.979
Ours	0.098	0.981	2.975	0.139	0.893	0.947	0.972

According to the method indicators calculated from the test results in Table 2, it can be seen that the designed feature adaptive fusion module and the network structure advantage, generalization and loss function effectiveness of the PVT v2transformer as the encoder backbone network.

IV Conclusion

In this paper, an encoder with a PVT v2 transformer as the backbone network is proposed, and a lightweight encoder with the adaptive fusion selection feature module is designed to estimate the depth of monocular images. The neural network is used to better extract global and local features and generate an Accurately estimated depth map. The encoder designed in this paper can adaptively select feature fusion in the extraction feature fusion stage to improve the estimated depth accuracy and introduce the mean square loss function to reduce the error. The effectiveness and generalization ability of the network is proved by the comparison of experimental results.

References

[1] Groenendijk R, Karaoglu S, Gevers T, et al. On the benefit of adversarial training for monocular depth estimation[J]. Computer Vision and Image Understanding, 2020, 190vol: 102848-50.

[2] Yang N, Stumberg L V, Wang R, et al. 2020. *D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry[C].*

2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle. 1278-1289.

[3] Fu H, Gong M, Wang C, et al. 2018 *Deep ordinal regression network for monocular depth estimation*[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City. 2002-2011.

[4] Eigen D, Puhrsch C, Fergus R. 2014. *Depth map prediction from a single image using a multi-scale deep network*[J]. Advances in neural information processing systems. 2283.

[5] Long J, Shelhamer E, Darrell T., 2015. *Fully convolutional networks for semantic segmentation*[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Boston. 3431-3440.

[6] Chen X, Chen X, Zha Z J. 2019. *Structure-aware residual pyramid network for monocular depth estimation[J]*. arXiv, vol: 1907.06023.

[7] Kim D, Ga W, Ahn P, et al. 2021. *Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth[J]*. arXiv preprint arXiv, vol: 2201.07436.

[8] Wang Q D, Wang Q K, Cheng K, Liu Z H. 2022. *Depth estimation of monocular images with enhanced edges [J].* Journal of Huazhong University of Science and Technology (Natural Science Edition), vol: 36-42.

[9] Wang W, Xie E, Li X, et al. 2022. *Pvt v2: Improved baselines with pyramid vision transformer[J]*. Computational Visual Media, vol: 415-424.

[10] Xu D, Yang G, Liu X M, et al. 2022. *Small sample image classification based on adaptive feature fusion and transformation[J]*. Computer Engineering and Applicationvol: 1-14

[11] Ishii, Yasunori and Takayoshi Yamashita. 2021. *CutDepth: Edge-aware Data Augmentation in Depth Estimation[J]*. ArXiv, vol: abs/2107.07684.

[12] Hu J, Ozay M, Zhang Y, et al. 2019. *Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps With Accurate Object Boundaries[C].* IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa Village. 1043-1051

[13] Ruan X G, Yan W J, Huang J, Guo P Y. 2022. *Monocular depth estimation method based on dual discriminator generation countermeasure network[J]*. Journal of Beijing University of Technology, vol: 928-934