Indoor Monocular Image Depth Estimation Based on Semantic Information of Tree-shaped ASPP Structure

Zhisong Bai
Changchun University of Science and Technology
Changchun, China
2579421789@qq.com

Chao Zhang *
Changchun University of Science and Technology
Changchun, China
* 454919258@qq.com

Cheng Han *
Changchun University of Science and Technology
Changchun, China
hancheng@cust.edu.cn

Linke Zhang
Changchun University of Science and Technology
Changchun, China
283646527@qq.com

2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA) | 978-1-6654-5911-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CVIDLICCEA56201.2022.9825336

***Abstract*—ASPP (Atrous Spatial Pooling Pyramid) has the advantage that it can expand the receptive field and extract multi-scale features without changing the image resolution. We introduce it into the depth estimation task to improve the problems of inaccurate depth estimation, blurred edges, and loss of depth information details on the current unsupervised depth estimation methods for indoor monocular images. However, the ASPP module does not consider the relationship between different pixel features, resulting in inaccurate extraction of scene features in the depth estimation task. Therefore, we propose a Tree-shaped ASPP structure for this drawback, combined with the SC-SfMLearner network using the NYUv2 dataset, adding the spatial semantic information pool formed by the ASPP tree structure between the encoder and decoder structures of the depth estimation network, which can not only expand the receptive field without losing resolution but also capture and fuse multi-scale context information, so that different pixels establish connections between features. The results show that, compared with the original method, the improved method has stronger network feature extraction ability, clearer contours of each target in the scene, more distinct layers, and more accurate depth estimation results.**

Keywords-unsupervised learning; depth estimation; semantic information; dilated convolution

I. INTRODUCTION

Scene depth estimation is a basic task in scene understanding. Depth information plays a crucial role in understanding the three-dimensional geometry and spatial relationship in the scene. The accuracy of many computer vision tasks requires accurate and effective depth information as the basis. Such as 3D reconstruction, vision-based automatic driving, SLAM, visual odometry, and other fields, at the same time, the accuracy of depth also reflects the computer's understanding of the scene.

The current depth information acquisition methods mainly include the direct acquisition method and the indirect acquisition method. The direct acquisition method is to directly acquire the depth through hardware equipment, but the required equipment is generally more expensive and has higher requirements for shooting, so the application range is small. The indirect acquisition method is to acquire depth information based on image processing. Traditional depth estimation methods such as Structure from Motion and Stereo Vision Matching are all based on the mutual correspondence of multi-view features, and the predicted depth map is sparse. Inferring depth information from a single image is a classic problem in computer vision, and it is also an ill-posed problem. It plays an indispensable role in augmented reality occlusion, illumination processing, and 3D reconstruction of scenes. Deep neural networks have developed rapidly in recent years and have shown excellent performance in image processing such as image classification[1], object detection[2], and semantic segmentation[3]. Monocular depth estimation based on deep learning has also been extensively studied and can be divided into supervised[4], unsupervised[5] and semi-supervised[6] depth estimation methods according to different training methods. The datasets on which supervised and semi-supervised depth estimation methods are trained to need to have ground-truth depths. However, it is difficult to obtain the true depth value of real scenes. Such datasets are difficult to produce and the number is small. The limitation of datasets makes it difficult for the algorithm to adapt to other scenarios, and the generalization ability is also weak.

Given the drawbacks of supervised and semi-supervised learning forms, unsupervised learning has received extensive attention because it does not require a large amount of measured ground truth data to participate. Zhou T H et al.[7] proposed a network overall architecture consisting of a depth estimation network (Depth CNN) and a camera pose estimation network (Pose CNN). The unlabeled monocular image sequence is used as the training set to realize the training of the monocular depth network and the camera pose estimation network, and the neural network is trained in an unsupervised manner for monocular depth estimation. Based on the depth and camera pose calculated by the neural network, it restores another image adjacent to an image in the picture sequence, and then calculates the pixel difference between the restored image and the real image, As the loss function for training, the network is finally converged by minimizing the loss function. This method of viewing synthesis provides ideas for later unsupervised monocular depth estimation networks, which are widely adopted.

II. SC-SFMLEARNER DEPTH ESTIMATION NETWORK STRUCTURE

The codec we use comes from the improved unsupervised monocular depth estimation algorithm SC-SfMLearner[8]. Garg et al.[10] first proposed to use color invariance to train a

978-1-6654-5911-2/22/$31.00 ©2022 IEEE

monocular depth estimation network on binocular images with known internal and external parameters of the camera. Based on the network proposed by Garg et al., Zhuo et al.[7] proposed a monocular depth estimation network SfMLearner trained on monocular video with camera intrinsic parameters. Based on SfMLearner, Bian et al.[9] proposed to use the geometric consistency loss as a constraint to estimate the depth consistency between the front and rear frames of the video, and then use this geometric consistency to detect moving objects and occluded areas. Removing these ill regions when calculating photometric loss can improve the performance of the algorithm, which is the SC-SfMLearner network.

This class of unsupervised algorithms works well with outdoor datasets, but it is difficult to train on indoor datasets. Bian et al. believed that the complex camera motion created an obstacle to training[8]. Finally, it is concluded through experiments that the key factor affecting the poor performance of unsupervised deep learning in indoor scenes is the motion of the camera in the training video, and the ideal data should be "no rotation + moderate translation". Therefore, Bian et al. processed the NYUv2 dataset and finally obtained a total of 67K ideal images for training.

III. Monocular Image Depth Estimation Network Fusing Semantic Information of Tree-shaped ASPP Structure

Since the current unsupervised monocular depth estimation network can achieve good results in outdoor scenes, but in indoor scenes, there are problems of inaccurate depth estimation, blurred edges, and loss of depth information details. We propose a monocular image depth estimation network that fuses the semantic information of Tree-shaped ASPP Structure.

A. Tree-shaped ASPP Structure

Chen et al.[11] proposed an ASPP structure based on dilated convolution in the Deeplabv2 network, and used four dilated convolutions with different sampling rates in the feature top map, which played a good role in semantic segmentation tasks. Deeplabv3[12] further improved the ASPP structure based on Deeplabv2 and proposed a parallel ASPP structure. However, the parallel ASPP structure will still lose some local information, and the information obtained by each layer is not connected and cannot be mutually dependent. To capture hierarchical context information and represent objects at multiple scales in complex scenes, we propose a Tree-shaped ASPP Structure.

The core of the Tree-shaped ASPP Structure is 4 dilated convolution modules with BN layers and ELU layers with different dilation rates, and a global average pooling layer. The Tree-shaped ASPP Structure takes the features extracted from the backbone network as input, and each dilated convolution module follows the expansion and stacking rules to efficiently encode multi-scale features. In each expansion step, the input is copied to two branches, one that preserves the current region features and the other that aggregates contextual information to explore the connections between the extracted features on a larger scale. At the same time, the output features of the current step are stacked with the previous features through concatenation. As shown in Figure. 1.

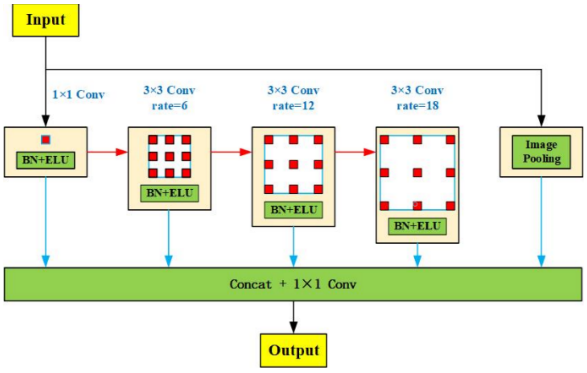

Figure 1. Tree-shaped ASPP Structure

The purpose of this design is that the Tree-shaped ASPP Structure can capture contextual information, represent multi-scale objects, and aggregate them so that the features learned from the previous steps can be re-studied in the subsequent steps so that the features of different layers are related.

B. Depth Estimation Network Structure

The network model we use is the SC-SfMLearner[9] network architecture, which includes a depth estimation network and a camera pose estimation network, and incorporates a Spatial Semantic Information Pool using a Tree-shaped ASPP Structure in the codec of the depth estimation network. As shown in Figure. 2.

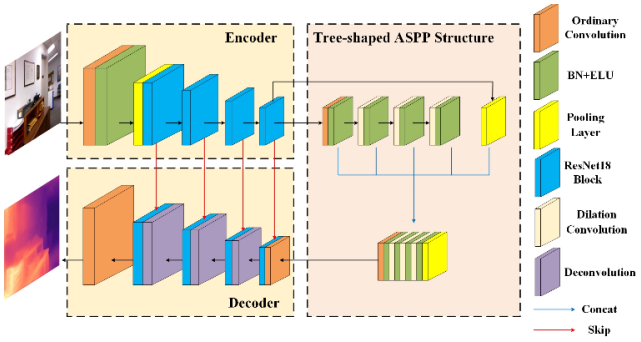

Figure 2. The depth estimation network model

The encoder of the network uses ResNet18 as the backbone network for feature extraction, which avoids the problems of feature degradation and low effectiveness. The decoding-side reverse convolutional neural network performs image reconstruction tasks, and there are skip connections from the encoder to the decoder to help the back-propagation of the gradient and speed up the training. Different from the traditional codec structure, the Tree-shaped ASPP Structure module is added in the middle of the codec of the network to improve the problem of edge blurring in the scene.

C. The Loss Function of The Model

The goal of the network is to train depth and pose CNNs from unlabeled video image sequences and constrain them to predict scale-consistent results. The overall objective function is as follows:

$$L = \alpha L_p^M + \beta L_s + \gamma L_{GC} \quad (1)$$

L_p^M represents the weighted photometric loss of the proposed mask M, L_s represents the smoothness loss, and L_{GC} represents the geometric consistency loss.

$$L_p^M = \frac{1}{|V|}\sum_{p\in V}\left(M(p)\cdot L_p(p)\right) \tag{2}$$

where M represents the self-discovery mask for handling moving objects and occlusions, and V represents the valid points where I_a is successfully projected onto the I_b image plane.

Since the photometric loss is not informative in low-texture or homogeneous regions of the scene, a smoothness loss is adopted before estimating the depth map, and we adopt the edge-aware smoothness loss used in[13]:

$$L_s = \sum_p\left(e^{-\nabla I_a(p)}\cdot\nabla D_a(p)\right)^2 \tag{3}$$

We enforce geometric consistency on the prediction results, specifically requiring the depth maps D_a and D_b of adjacent frames to conform to the same 3D scene structure and minimizing their differences, not only encouraging geometric consistency between samples in batches, and transfer the identity to the entire sequence. Under this constraint, the geometric consistency loss is:

$$L_{GC} = \frac{1}{|V|}\sum_{p\in V}\frac{|D_b^a(p)-D_b'(p)|}{D_b^a(p)+D_b'(p)} \tag{4}$$

This minimizes the geometric distance of predicted depths between each consecutive pair and enforces its scale consistency. Through training, consistency can be propagated to the entire video sequence. Due to the close connection between camera motion and depth prediction, the camera pose estimation network can finally predict trajectories that are globally scale-consistent.

IV. Network Training and Test Evaluation

We performed three sets of ablation experiments to verify the effectiveness of our proposed network model. The invariant factors are consistent, and the experiments are carried out on the same computer, in the same environment, and with the same parameters.

A. Model Training Parameters

The graphics card used in the experiment is NVIDIA RTX 3090. We ado the processed NYUv2 dataset[8] for training and testing. The training set includes 67K images and their corresponding ground-truth depth maps, each with a resolution of 320×256. The test set includes 1449 monocular images and their corresponding ground-truth depth maps, each with a resolution of 640 × 480. Our network model is based on the PyTorch framework and optimized using the Adam optimizer with parameters set to $\beta_1 = 0.9$、$\beta_2 = 0.999$. During training, the initial learning rate of the experimental setting is 0.0001, the batch size is 16, and a total of 50 epochs are trained.

B. Error Evaluation Index

$$AbsRel = \frac{1}{|N|}\sum_{d\in N}\frac{|d-d^*|}{d^*} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{|N|}\sum_{d\in N}\|d-d^*\|^2} \tag{6}$$

$$lg = \frac{1}{|N|}\sum_{d\in N}|\lg d - \lg d^*| \tag{7}$$

$$Thr = \%of\ d_i\ such\ that\ max\left(\frac{d_i}{d_i^*},\frac{d_i^*}{d_i}\right) = \delta < thr \tag{8}$$

Among them, d is the depth map predicted by the network, d^* is the ground truth data, thr is a given threshold, and its parameters are set to $\delta_1 < 1.25$, $\delta_2 < 1.25^2$, and $\delta_3 < 1.25^3$.

C. Evaluation of Quantitative Indicators

In this experiment, an ablation experiment is designed to verify the effectiveness of the core module proposed in this paper. The original network model SC-SfMLearner, adding the Spatial Semantic Information Pool with ASPP as the core, and adding the Spatial Semantic Information Pool with the Tree-shaped ASPP Structure as the core are respectively evaluated by three sets of comparative experiments. The evaluation results are shown in Table 1.

TABLE I. Evaluation and Comparison of Quantitative Indicators of Ablation Experiments

Ablation Analysis	Comparative Experiment		
	Original Experiment	***With ASPP***	***With Tree-shaped ASPP***
AbsRelpt↓	0.148	0.148	0.146
Log10↓	0.062	0.062	0.062
RMSE↓	0.543	0.539	0.535
$\delta_1 < 1.25$↑	0.803	0.804	0.806
$\delta_2 < 1.25^2$↑	0.949	0.948	0.951
$\delta_3 < 1.25^3$↑	0.985	0.985	0.986
Params	14.15M	31.66M	31.66M
Model Size	59.5MB	95.1MB	132.9MB
Training Time	13h45min	15h	15h
Test Time	52s	55s	55s

It can be seen from Table 1 that: Comparing the three groups of experiments, the experimental error evaluation index with the tree-shaped ASPP structure performs the best. Comparing the original model with the experiment of adding the original ASPP structure, it can be seen that the introduction of the ASPP structure in the network will become better. Comparing the experiment of adding the original ASPP structure and the experiment of adding the Tree-shaped ASPP Structure, it can be seen that the Tree-shaped ASPP Structure has an improvement effect on the ASPP structure.

D. Comparative Analysis of Visualization Results

The essential task of depth estimation is to generate reasonable and reliable depth maps. Figure. 3 is a visual comparative analysis of the original experiment, the experiment adding the ASPP structure, and the experiment adding the Tree-shaped ASPP Structure on the same image, including the depth map, the original RGB map, and the ground truth value.

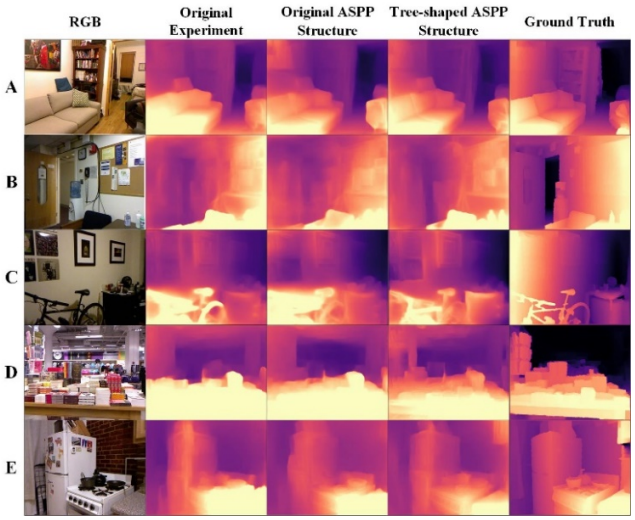

Figure 3. Visual comparison of ablation experiment results

As shown in Figure. 3, in the indoor complex scene, the original experiment performed the worst, and the experiment with the Tree-shaped ASPP Structure was the best. The result of the original experiment lacks the depth level, the edge position of the object is not clear, there is a phenomenon of deep connection, and the depth estimation of some positions has a big error. The experiment of introducing the ASPP module is slightly better than the original experiment, but the experimental result of adding the Tree-shaped ASPP Structure module is even better, the edge of the object is clearer, the depth hierarchy can be seen, and the depth estimation is more accurate.

V. CONCLUSION

Aiming at the existing problems of unsupervised monocular depth estimation network, we propose a network model combining semantic information modules based on the existing network. We make structural changes to the commonly used ASPP modules in semantic segmentation tasks and propose a Tree-shaped ASPP Structure. We add a Spatial Semantic Information Pool with a Tree-shaped ASPP Structure as the core between the traditional codec structures to expand the receptive field of convolution, capture contextual information, and avoid the loss of detailed information. The extracted features are integrated to make the features of different layers connected, and the final monocular depth estimation results have achieved excellent results.

ACKNOWLEDGMENT

The authors feel like showing the sincerest as well as grandest gratitude to the Changchun University of Science and Technology. This project received great support from the Science and Technology Development Program of Jilin Province, China (No. 20200403188SF), the National Natural Science Foundation of China (No. 61702051).

REFERENCES

[1] Zhang F, Zhu X, Ye M. Fast human pose estimation [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019:3517-3526.

[2] Pang J, Chen K, Shi J, et al. Libra R-CNN: Towards balanced learning for object detection [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019:821-830.

[3] Lyu H, Fu H, Hu X, et al. ESNet: Edge-based segmentation network for real-time semantic segmentation in traffic scenes [C]// 2019 IEEE International Conference on Image Processing. Taipei: IEEE, 2019:1855-1859.

[4] Kendall A, Martirosyan H, Dasgupta S, et al. End-to-end learning of geometry and context for deep stereo regression [C]// Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017:66-75.

[5] Mahjourian R, Wicke M, Angelova A. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018:5667-5675.

[6] Kuznietsov Y, Stuckler J, Leibe B. Semi-supervised deep learning for monocular depth map prediction [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017:6647-6655.

[7] Zhou T H, Matthew B, Noah S, et al. Unsupervised learning of depth and ego-motion from video [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6612-6619.

[8] Bian J W, Zhan H, Wang H, et al. Unsupervised Depth Learning in Challenging Indoor Video: Weak Rectification to Rescue [J]. CoRR, 2020, abs/2006.02708.

[9] Bian J W, Li Z C, Wang N Y, et al. Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video [J]. CoRR, 2019, abs/1908.10553.

[10] Garg R, Bg V K, Carneiro G, et al. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue [J]. Lecture Notes in Computer Science, 2016, vol 9912.

[11] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, vol.40, no.4: 834-848.

[12] Chen L C, Papandreou G, Schroff F, et al. Rethinking Atrous Convolution for Semantic Image Segmentation [J]. CoRR, 2017, abs/1706.05587.

[13] Ranjan A, Jampani V, Kim K, et al. Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019:12232-12241.