Hindawi
Journal of Electrical and Computer Engineering
Volume 2019, Article ID 9340129, 10 pages
https://doi.org/10.1155/2019/9340129

Research Article

Traffic Scene Depth Analysis Based on Depthwise Separable Convolutional Neural Network

Jianzhong Yuan,[1] Wujie Zhou ,[1,2] Sijia Lv,[1] and Yuzhen Chen[1]

[1]*School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China*
[2]*Institute of Information and Communication Engineering, Zhejiang University, Hangzhou 310027, China*

Correspondence should be addressed to Wujie Zhou; wujiezhou@163.com

Received 4 March 2019; Revised 29 May 2019; Accepted 4 June 2019; Published 19 June 2019

Academic Editor: Cesare F. Valenti

In order to obtain the distances between the surrounding objects and the vehicle in the traffic scene in front of the vehicle, a monocular visual depth estimation method based on the depthwise separable convolutional neural network is proposed in this study. First, features containing shallow depth information were extracted from the RGB images using the convolution layers and maximum pooling layers. Subsampling operations were also performed on these images. Subsequently, features containing advanced depth information were extracted using a block based on an ensemble of convolution layers and a block based on depth separable convolution layers. The output from all different blocks is combined afterwards. Finally, transposed convolution layers were used for upsampling the feature maps to the same size with the original RGB image. During the upsampling process, skip connections were used to merge the features containing shallow depth information that was obtained from the convolution operation through the depthwise separable convolution layers. The depthwise separable convolution layers can provide more accurate depth information features for estimating the monocular visual depth. At the same time, they require reduced computational cost and fewer parameter numbers while providing a similar level (or slightly better) computing performance. Integrating multiple simple convolutions into a block not only increases the overall depth of the neural network but also enables a more accurate extraction of the advanced features in the neural network. Combining the output from multiple blocks can prevent the loss of features containing important depth information. The testing results show that the depthwise separable convolutional neural network provides a superior performance than the other monocular visual depth estimation methods. Therefore, applying depthwise separable convolution layers in the neural network is a more effective and accurate approach for estimating the visual depth.

1. Introduction

Automobiles have become an indispensable means of transportation for peoplenowadays. Development of advanced automobile has always been an important task for the society. Estimating the monocular visual depth in the images obtained in front of the vehicle can provide great assistance to the driving experience and, furthermore, ensure that the driving is safe. A mature and accurate depth estimation technology in a traffic scene can effectively ensure that the drivers and passengers remain safe on road [1–3].

Monocular depth estimation is the key to the reconstruction and understanding of the scene. The initial use of monocular depth estimation was to use statistically significant monocular clues or features, such as perspective and texture information, object size, object position, and occlusion [4–6]. Recently, some works use convolutional neural network-based models to significantly improve monocular depth estimation performance [7–11], indicating that depth features are superior to manual features. These methods resolve the monocular depth estimation problem by learning the convolutional neural network to estimate the sequential depth maps. Since this problem is a standard regression problem, the mean square error (MSE) of logarithmic space or its variant is usually used as the loss function.

The application of the convolutional neural network based on the deep learning method in depth estimation has promoted the rapid development of depth estimation technology. Various depth estimation networks have been proposed to improve the accuracy of depth estimation results, which makes the application of the visual depth estimation technology more and more extensive. The proposed and innovative methods can not only improve the accuracy of depth estimation but also enable the predicted depth map to be applied to other computer vision tasks and improve its results.

In this study, we constructed a block based on depthwise separable convolution layers and combined it with the block used in ResNet for extracting the advanced feature information. Meanwhile, the low-level features obtained from the convolution operation through the depthwise separable convolution layers are merged using the skip-connection technique during the decoding process. The testing results show that the neural network model developed in this study provides more accurate depth estimation in a monocular image compared to other available methods. The following innovations are featured in this study: (1) The application of depthwise separable convolution layer can reduce the computational cost and parameter number while maintaining a similar (or slightly better) performance. This approach can capture the depth information features more accurately, which further improve the accuracy of the final predicted depth map. (2) The block structure used in ResNet is referred in this study. Based on the depthwise separable convolution layer, a block structure similar to that used in ResNet is established in this study and used in conjunction with the block in ResNet to constitute part of the encoder of the neural network model. This method allows the outputs from all different blocks to be merged together without changing the size of the characteristic map. Therefore, a sufficient depth required for extracting abundant feature information is ensured in the model, which also makes the model framework more accurate. (3) The characteristics of skip connection not only allow us to piece together the missing edge information associated with the advanced features but also further provide the edge depth information through the depthwise separable convolution. The information contributes to a more accurate output from the final model.

2. Related Work

Before the application of deep learning methods, the task of estimating monocular depth in the field of computer vision has been accomplished by manual extraction of the key features traditionally. The earlier studies were primarily focused on the depth estimation of stereo images. In these studies, the depths are calculated using a geometric algorithm [2, 3] based on the 3D points determined by image triangulation. Since then, different manual methods for expressing the monocular depth feature have been proposed [4, 5, 12–19]. Since the features extracted by manual approaches can only capture local information, probabilistic graphical models such as Markov random fields (MRFs) are usually constructed based on these features in order to include the long distance and global information [4, 20, 21]. Another successful method to use global information is the DepthTransfer method [6]. Such a method uses the GIST global scene feature [22] to search for candidate images that are "similar" to the input image in a database containing RGB-D images.

In recent years, deep learning-based depth estimation techniques have been proposed continuously by different studies. Some depth estimation methods using deep convolutional neural networks have been widely studied. Eigen et al. [23] proposed a method which combines two depth estimation networks: a coarse network that predicts the global depth distribution and a fine network which refines the detail of the local depth map. Eigen and Fergus [7] further extended this study to a three-level network structure and performed surface normal estimation and semantic label estimation, as well as depth estimation. Roy and Todorovic [8] integrated a relatively shallow CNN into a forest regression method. Laina et al. [9] designed a depth estimation network [24] based on the ResNet architecture. They proposed a top project structure to improve the resolution of the depth map. In addition, they proposed the concept of Huber loss in the network training which combines the Euclidean function and L1 function. Mancini et al. [25] implemented a long short-term memory layer in the neural network model. This method alleviates some inherent limitations of monocular vision such as global scale estimation in the sequential image stream and reduces the computational cost. Xie et al. [26] employed the skip-connection strategy to combine the low spatial resolution depth maps in deeper layers with high spatial resolution depth maps in lower layers.

In order to improve the progress of depth estimation, a conditional random field (CRF) is added to the neural network. Yin et al. [27] applied depth estimation in scale recovery and combined it with CRF, which further improves the accuracy of the results. Wang et al. [10] trained a CNN for joint depth estimation and semantic segmentation. They further employed a CRF model to improve the prediction results from the CNN. Xu et al. [28] integrated the side-output maps from the CNN using multiple consecutive CRFs. This method involves extracting features from each single layer of the network and combining the features to estimate the depth map.

There are some other methods that are used for the depth estimation. Li et al. [29] proposed a dual flow CNN with high training speed for predicting depth and depth gradient. The information was combined to yield an accurate and detailed depth map. Li et al. [30] treated monocular depth estimation as a heavy multicategory marking task. They combined the side outputs from an expanded convolutional neural network following a hierarchical manner and performed depth estimation using multiscale depth cues. In addition, they recommended to use soft weighting and reasoning instead of hardware reasoning to convert the discrete depth values into continuous depth values. Chakrabarti et al. [31] predicted the derivatives of the depth at different orders in a probabilistic manner and estimated the depth map through a localization process. Ummenhofer et al. [32] trained a

convolutional network that calculates depth and camera motion from a continuous, unconstrained image end-to-end. Lee et al. [33] proposed a deep learning algorithm for depth estimation in a monocular image based on the Fourier analysis in the frequency domain. Fu et al. [34] proposed a deep ordinal regression network (DORN) to improve the computation efficiency. In their method, the true depths are first discretized into a number of subintervals using a spacing-increasing discretization strategy. Subsequently, a pixelwise ordinal loss function is designed to simulate the ordinal relationships of these depth subintervals. The different methods proposed in the literature studies have greatly promoted the development of depth estimation techniques. However, some of these methods ignored the importance of the depth of the neural network for extracting feature information. Some other methods may have sufficient neural network depth but fail to include the shallow feature information which results in reduced accuracy of the overall depth estimation result.

Deep residual neural network (ResNet) has achieved great success in computer vision applications. Furthermore, Chen et al. [35] have successfully applied depthwise separable convolution layers in the field of semantic segmentation computer vision. In light of these achievements, we proposed a new approach to estimate visual depth in this study. We first constructed a block by combining the skip connections with the conventional convolution layers based on the depth separable convolution layer (SeparableConv2D layer). Subsequently, this block is combined with the block in ResNet for extracting features containing depth information.

When dealing with depth estimation in a monocular image, the use of block can help build a neural network with sufficient depth for extracting features containing greater depth information with higher accuracy. Furthermore, using block with residual characteristics can effectively resolve the vanishing gradient problem encountered when stacking more neural network layers. In depthwise separable convolution, the standard convolution is firstly decomposed into a depthwise convolution and then a pointwise convolution. This approach greatly reduces the computational complexity. Specifically, a spatial convolution is performed on each input channel independently in the depthwise convolution, while the outputs from the depthwise convolutions are combined by pointwise convolution. In addition, depthwise separable convolution significantly reduces the computational complexity of the proposed model while maintaining a similar (or better) performance.

After Laina et al. [9] successfully applied ResNet to the depth estimation task, we find the efficient performance of the residual block in ResNet [24] and modify it to build a more suitable residual block for depth estimation. The depth separable convolution layer has been used successfully by Chen et al. [35], which proved that the features with high accuracy depth information can be extracted by the depth separable convolution layer. It is very important for depth estimation task. Therefore, we reference the advantage of residual properties to build a new block, which is based on the depth separable convolution layer. Furthermore, Zhou et al. [36] and Yang et al. [37] connected multilayer feature information by skip connections that are of great help for the final results. Therefore, we use skip connections in our neural network architecture.

3. Proposed Model Method

3.1. Network Architecture Framework. An end-to-end learning framework is used in this study to accomplish the computer vision task of monocular depth estimation. Such a framework involved learning the direct mapping from a colorful image to the corresponding depth map. In recent years, a considerable amount of studies has shown that the depth of the CNN architecture is very important to the performance of the network. However, simply stacking more layers in the current CNN architecture does not necessarily guarantee an improvement and may instead leads to a drastic reduction in the performance. This is because simply stacking layers in the CNN can cause vanishing gradient problem which prevents feature fusion in the neural network when training starts. Residual neural network has been developed to tackle the vanishing gradient issue and achieved huge success in major computer vision fields. One of the most representative examples, ResNet, has been very popular in the field of computer vision tasks.

Two types of residual blocks in ResNet, conv_block and identity_block, are used as references for this study. Following their structures, two similar blocks are established in this study using depthwise separable convolution: the sep_block and ap_block. Furthermore, the blocks sharing a similar structure such as conv_block and sep_block or identity_block and ap_block are used together to establish the neural network model in this paper. Different from the original conv_block and identity_block, all the blocks used in this study exhibit a convolution timestep of 1. Meanwhile, the size and number of filters used in the conventional convolution layers and the depthwise separable convolution layers are same for all blocks. To distinguish from the residual block used in the original ResNet model, the blocks constructed in our study are named conv_block and identity_block. Figure 1 shows the main structure of the neural network model constructed in this study.

The extraction of shallow feature information and downsampling of the images are performed using three conventional convolution layers and three maximum pooling layers in this network system. The advanced features are extracted by using a combination of conv_block and sep_block plus a combination of identity_block and ap_block. There are two ways to use these blocks in this study: the first way is to use conv_block together with sep_block, while the other way is to use identity_block together with ap_block. These two approaches are used separately. In other words, two types of neural network model are constructed in this study, but the main structure of these models is still represented by Figure 1. When merging the outputs from six blocks using the concatenate layer, the feature map is upsampled using a transposed convolution layer after going through a convolution layer. This process is repeated until the feature map is restored to

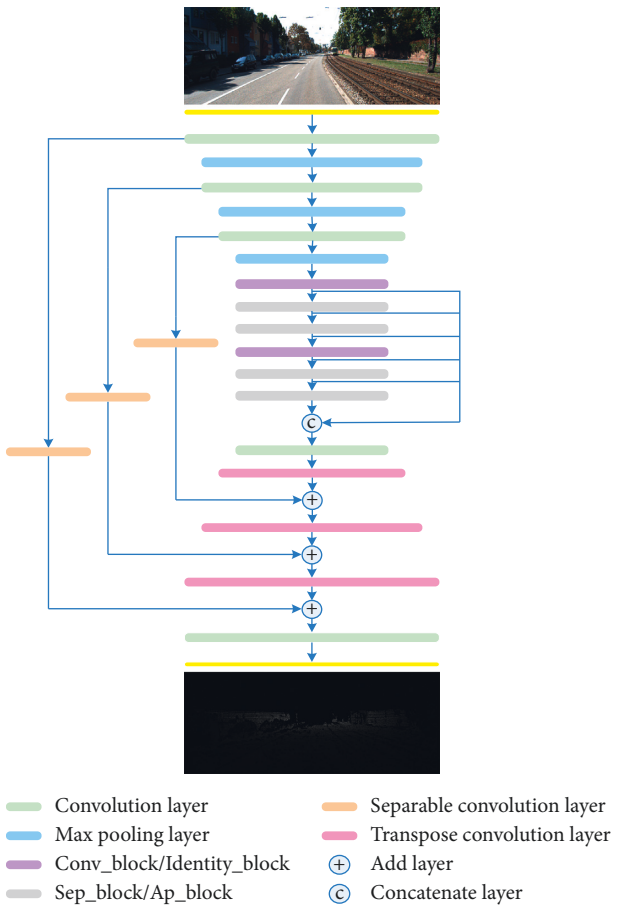

Figure 1: The neural network structure. There are only two ways to use conv_block/identity_block and sep_block/ap_block. One is to use conv_block combined with sep_block, and the other is to use identity_block combined with ap_block.

the same size with the original image for yielding the final depth map predicted by the neural network. Meanwhile, considering the advantage of residual characteristics, the outputs from three convolution layers used for extracting low-level features are processed by three depthwise separable convolution layers during the transposition convolution operation. The outputs from the three depthwise separable convolution layers are subsequently fused to the output from three transposed convolution layers of the same size using three Add layers, respectively.

It should be noted that a batch normalization layer and a ReLu layer are implemented after every single convolution layer and transposed convolution layer. These additional layers are not shown in the figure for simplifying the model structure schematic. The detailed composition of the four blocks will be discussed in Sections 3.2 and 3.3.

3.2. Deep Residual Network Block. Using deep residual network blocks can not only extend the depth of the neural network and improve its performance but also prevent the network performance from degradation with increasing network depth.

Two residual network blocks in ResNet, conv_block and identity_block, are modified slightly and used in this study as conv_block and identity_block. The detailed structure maps of these blocks are shown in Figure 2.

The identity_block shown in Figure 2(a) is defined as

$$\mathbf{y} = F(\mathbf{x}, \{W_i\}) + \mathbf{x}, \tag{1}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the input and output matrices of the stacked layer, respectively. The function $F(\mathbf{x}, \{W_i\})$ is the residual mapping that needs to be learnt. Since skip connections are performed by adding the elements, $\mathbf{x}$ and F must share the same size.

The conv-block shown in Figure 2(b) is defined as

$$\mathbf{y} = F(\mathbf{x}, \{W_i\}) + W_i\mathbf{x}. \tag{2}$$

Similar to equation (1), the size of $\mathbf{x}$ and F must be equal.

3.3. Depthwise Separable Convolutional Network Block. Depthwise separable convolution layers can reduce the computational cost and number of parameters while maintaining a similar (or slightly better) performance. The network blocks sep_block and ap_block are constructed based on depthwise separable convolution layers but following the structures of conv_block and identity_block. The detailed structure maps of these two blocks are shown in Figure 3.

In depthwise separable convolution, the standard convolution is first decomposed into a depthwise convolution and then into a pointwise convolution (i.e., the convolution layer on the left side in sep_block). Specifically, a spatial convolution is first performed on each input channel independently in the depthwise convolution, and pointwise convolution is used subsequently to merge the outputs from the depthwise convolutions. The deep convolution supports setting up an expansion ratio. The depthwise separable convolution layers set with an expansion ratio are called as atrous separable convolution layers. This type of convolution layer is employed in one of the neural network models established in this study. Specifically, only the second depthwise separable convolution layer among the three depthwise separable convolution layers located on the right side in sep_block is an atrous separable convolution layer. The detailed testing procedure and results will be discussed in the testing section.

3.4. Using of Skip Connection. Skip connection is usually used in residual networks and is an identity mapping method first proposed by ResNet. In ResNet, a residual network structure called residual network is introduced, which is different from the ordinary CNN in that it connects an additional transfer line directly from the input source to the output source, which is a kind of identity mapping, used for residual calculation. This is called shortcut connection, and it is also called skip connection. The effect is to prevent gradient dispersion and degradation problems caused by the increase of network layers.

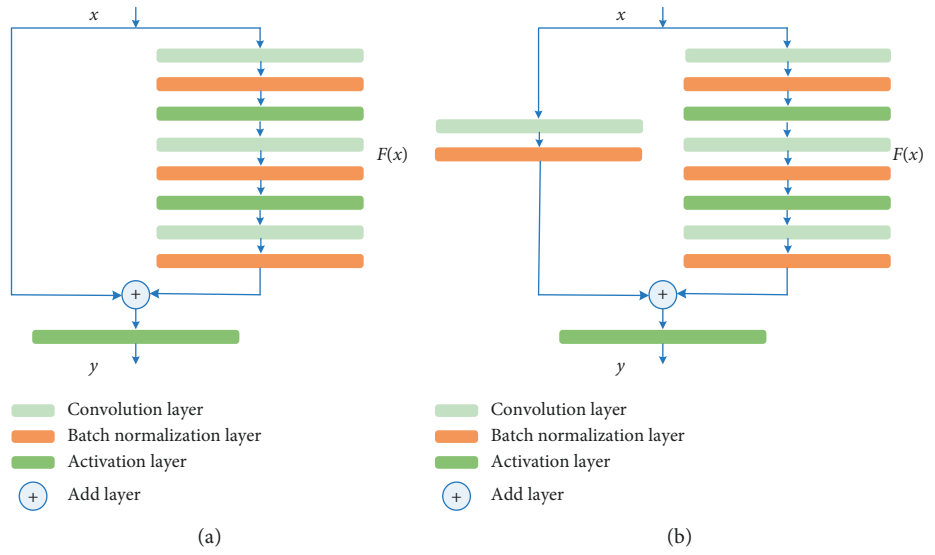

Figure 2: Two types of blocks. (a) identity_block. (b) conv_block.

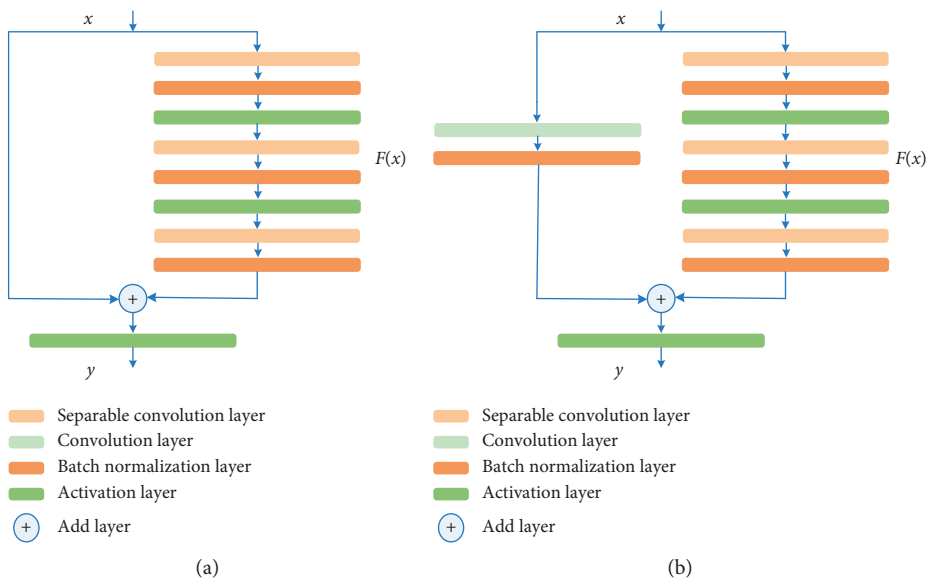

Figure 3: Two types of blocks. (a) ap_block. (b) sep_block.

In this article, we mainly use two skip-connection methods. One is the summation based on elements, such as the Add layer in Figure 1, and the other is the splicing based on elements, such as the concatenate layer in Figure 1. We use the concatenate layer to concatenate all the features of block output to ensure the full use of these advanced feature information. In addition, after multilayer feature extraction, many edge contour information of the object in the original image has been lost in the advanced features, which is very unfavorable to the final depth prediction result. Therefore, we use the Add layer to fuse the low-level features with rich contour information in the initial stage of coding with the high-level features in the decoding stage in the way of element addition so as to improve the accuracy of the prediction results of the whole network.

4. Experiments and Analysis of Results

4.1. Data Set Introduction. The image data set used for network training and testing in this paper is the outdoor depth estimation data set—KITTI data set [38]. It is also the largest public outdoor data set in the field of depth estimation. The data set contains real image data collected from scenes such as city streets, country roads, and highways. Each image contains a variety of contents including multiple cars and pedestrians. Different levels of occlusion and truncation are also present in the images. Due to the vast amount of data included, this data set can be used for training complex deep learning models that are used for depth analysis or depth prediction in a single image. Due to the numerous outdoor scenes included, the KITTI data set

has been used in many computer vision tasks and becomes a recognized standard for testing network models.

The data set used in this study contains two parts: the training data set and the testing data set. The training data set includes 4286 original colored images and the corresponding depth labels. It is primarily used for training the neural network model until the optimum weights are obtained. The testing data set includes 343 colored images and the corresponding depth labels. The testing data set is used for testing the accuracy of the prediction from the neural network model loaded with the optimum weights. It is further used for comparing the prediction from the neural network model with those given by other methods. During both the training and testing stage, all the images are imported into the network for extracting the required information. The final depth estimation map given by the prediction has the same size with the original colored images.

4.2. Training Method. Both the training and testing experiments of the deep neural network model are performed on the Keras learning framework, which is based on the TensorFlow backend. The computer used for performing these experiments is equipped with a GTX 1080 Ti GPU possessing 11 GB memory. The Adam algorithm [39] is used as the optimization method during the training process, and the learning rate is set at 0.01. The batch size used for neural network training is 4.

The neural network is constructed using an end-to-end deep learning framework. To be used as input to the neural network, the images from the data set must have the same size. However, the size of the images in the data set provided by KITTI is not completely consistent. Therefore, before training the neural network model, the size of the image needs to be processed in a unified way. After extracting the key features in the neural network, a predicted depth map with the same size will be generated from the input RGB image. In the process of model training, mean-squared error (MSE) is used as a loss function to obtain the optimal weight $\mathbf{W}^{\text{best}}$ in the training stage. The calculation formula of MSE is as follows:

$$\frac{1}{T}\sum_{i=1}^{T}\left(\mathbf{Y}_n - \overline{\mathbf{Y}}_n\right)^2. \tag{3}$$

Here, $\mathbf{Y}_n$ is the depth label of the nth original image in the data set, $\overline{\mathbf{Y}}_n$ is the predicted depth map obtained by processing the nth original image in our neural network model, and T is the training data set.

4.3. Experimental Methods and Comparison of the Results. Four different experimental methods of depthwise separable convolution are used in this study for depth estimation. Different results are generated from these methods. Based on the type of block used, conv_block or identity_block, in conjunction, these four methods can be classified into two categories. The first category is associated with sep_block which is constructed based on depthwise separable convolution and follows a similar structure with conv_block. This block is mainly composed of separable convolution layers in three depth directions, a conventional convolution layer, ReLu layers, and (batch) normalization layers. The detailed structure of sep_block is shown in Figure 3(b). These two models share a same neural network framework as shown in Figure 1. Within this category, we discussed how the experimental results will be affected by the use of atrous separable convolution in conv_block and sep_block. In the first experimental method, the neural network is constructed by combining the original conv_block and sep_block directly without the use of atrous separable convolution. The experimental results "Proposed1" obtained using this method are superior to those obtained by other classical depth estimation methods. A comparison of these results is shown in Table 1. In the second experimental method, an expansion ratio is set in the second conventional convolution layer/depth separable convolution layer among the three conventional convolution layers/depth separable convolution layers located on the right side in conv_block and sep_block. As shown in Figure 1, the expansion ratios of the first conv_block, the first sep_block, and the second sep_block are set to 1; the expansion ratios of the second conv_block and the third sep_block are set to 2; and the expansion ratio of the last sep_block is set to 4. The final experimental results "Proposed2" are shown in Table 1. Although the second method, compared to the first method, only yields a minor enhancement in the first two major indicators, it provides a significant general improvement compared to the other classical depth estimation methods.

The other two experimental methods used in this study belong to the second category. These methods involve the use of identity_block and ap_block together for performing depth estimation tests. These two blocks have a similar structure consisting of three conventional convolution layers/depthwise separable convolution layers, ReLu layer, and (batch) normalization layers. The detailed structure is shown in Figures 2(a) and 3(a). The neural networks established in the second category experiments are slightly different from those shown in Figure 1. Specifically, we added a conventional convolution layer before the first identity_block, the second identity_block, and the fourth ap_block in Figure 1. These three conventional convolution layers have the same number of filters with the three blocks. The experimental method in this category is the same with the first experimental method included in the first category, where the identify_block and ap_block are used together for performing the prediction. Table 1 shows the experimental results "Proposed3" obtained by this method which is superior to those obtained by other classical depth estimation methods.

4.4. Analysis of Experimental Results. When evaluating and comparing the experimental results, the evaluation index is obtained by comparing the predicted depth map with the depth label of the original image. In this study, six objective parameters which are commonly used for assessing monocular visual depth prediction are selected as the evaluation indices. The parameters include the following.

Table 1: Depth estimation results on the KITTI data set.

Method	Accuracy (higher is better)			Error (lower is better)		
	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	rms	log_rms	log10
Xie et al. [26]	0.488	0.947	0.972	2.6440	0.272	0.167
Laina et al. [9]	0.674	0.943	0.972	2.4618	0.243	0.126
Yin et al. [27]	0.640	0.947	**0.979**	2.5193	0.247	0.134
Dimitrievski et al. [40]	0.634	0.916	0.945	2.8246	0.305	0.127
Mancini et al. [25]	0.566	0.945	0.970	2.6507	0.264	0.145
Proposed1	**0.693**	**0.949**	0.975	**2.3993**	**0.232**	**0.117**
Proposed2	**0.691**	**0.949**	0.976	**2.3676**	**0.231**	**0.117**
Proposed3	**0.699**	**0.949**	0.975	**2.3786**	**0.232**	**0.116**

Root mean squared error (rms):

$$\sqrt{\frac{1}{T}\sum_{n\in T}\left(\mathbf{Y}_n - \overline{\mathbf{Y}}_n\right)^2}. \tag{4}$$

Logarithmic root mean squared error (log_rms):

$$\sqrt{\frac{1}{T}\sum_{n\in T}\left\|\log \mathbf{Y}_n - \log \overline{\mathbf{Y}}_n\right\|^2}. \tag{5}$$

Average log10 error (log10):

$$\frac{1}{T}\sum_{n\in T}\left|\log \mathbf{Y}_n - \log \overline{\mathbf{Y}}_n\right|. \tag{6}$$

Threshold accuracy (thr):

$$\max\left(\frac{\mathbf{Y}_n}{\overline{\mathbf{Y}}_n}, \frac{\overline{\mathbf{Y}}_n}{\mathbf{Y}_n}\right) = \delta < \text{thr}. \tag{7}$$

Here, $\mathbf{Y}_n$ is the depth label of the nth original image in the data set, $\overline{\mathbf{Y}}_n$ is the predicted depth map obtained by processing the nth original image in our neural network model, and T is the testing data set.

Table 1 shows the comparison of the depth estimation results obtained with the KITTI data set using the neural network proposed in this study and other classical models. A smaller error indicator value and a larger accuracy indicator value in Table 1 represent a better estimation performance. The data from our study which are superior to those from other classical studies are in bold. When testing the performances of the neural network model proposed in this study and other neural network models implemented for comparison, no data preprocessing step is required and involved. It is found that, for monocular visual depth estimation problem, the neural network model proposed in this study can provide much better results with a higher depth estimation accuracy compared to the other models.

The introduction of deep learning has brought rapid development in the field of monocular visual depth estimation as well as an increasingly higher estimation accuracy. However, the continuous expansion and deepening of the deep learning framework will result in increasing depth of the neural network. Following this trend, a larger and larger amount of data will be required for training and testing the neural network. However, the increasing depth of the neural network can cause disconnection of the information between the early and latter session of the network, loss of important feature information, and vanishing gradient problem during the training process. All these phenomena will lead to failure of the entire model. Meanwhile, an extremely deep neural network model contains a huge number of parameters itself. These parameters impose a very high demand on the hardware performance of the computer. Besides, training these parameters will also consume a considerable amount of time. These concerns are the bottlenecks of many deep neural networks. The introduction of residual networks can well avoid the vanishing gradient issue which provides great help to the continuous stacking of deep neural network layers. In addition, depthwise separable convolution can significantly reduce the computational cost and number of parameters without sacrificing the network performance and, furthermore, helping the network extract more features containing depth information.

Taking advantage of the residual characteristic, we established a neural network model using blocks in this study. This not only ensures a sufficient depth of the network but also allows us to apply this idea to the entire network. We first extracted the shallow-layer information using the depthwise separable convolution layers and further merged them with the advanced features. This approach fully preserves the feature information and allows us to obtain more accurate depth information features. At the same time, the features are extracted using the block without changing the size of the feature map. In addition, the outputs from multiple blocks are used in series for generating the features. These practices allow the neural network developed in this study to extract plentiful accurate advanced features. Moreover, since the blocks used to extract the main advanced features are composed of depthwise separable convolution layers, the computational cost and parameter amount of this network are much smaller than those of neural networks that extract features only using the two blocks in ResNet. The comparison shown in Table 1 demonstrates that the neural network model developed in this study has a superior performance compared to many other models.

Our neural network model did not yield a better performance indicator on $\delta < 1.25^3$ than the method proposed by Yin et al. This is because a conditional random field (CRF) is implemented in their method. However, we have performed an extensive number of studies and found that the increase in three performance indicators brought by the CRF is accompanied by the increase in three error performance

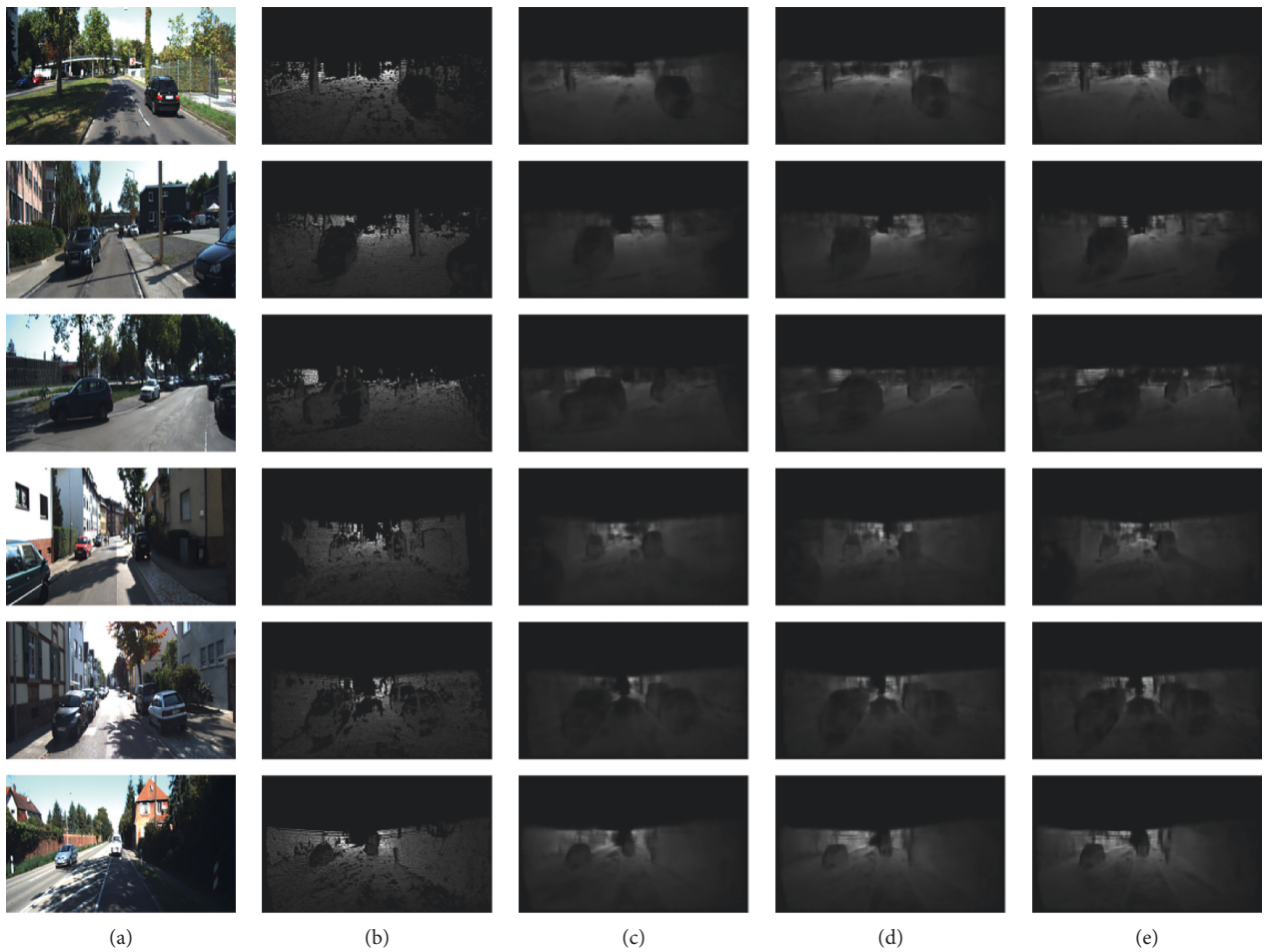

Figure 4: Experimental results chart. (a) RGB image, (b) ground truth, (c) depth prediction map of Proposed1, (d) depth prediction map of Proposed2, and (e) depth prediction map of Proposed3.

indicators. Therefore, employing the CRF in the model cannot provide an improvement to every single performance indicator. While one performance indicator from our method is inferior to that in another study, our method can still provide significant improvements in the other performance indicators, which suggests that the network model developed in this study has a general superior performance than the other network models.

The experimental results obtained in this study are shown in Figure 4. Figures 4(a) and 4(b) show the original RGB image and the depth label (ground t), respectively. Figures 4(c)–4(e) show the depth prediction map obtained using the experimental methods "Proposed1," "Proposed2," and "Proposed3," respectively.

5. Conclusion

It is very important to perceive the traffic environment and use the objects detected in the traffic scene to provide assistance to vehicles for traveling on road, in particular self-driving. The advancement made in the field of computer vision task for monocular depth estimation can provide significant help to the autonomous vehicle technology and ensure a better driving safety on road. In this study, we estimated the depth map using the images captured in front of the vehicle. These depth maps can provide important information on the traffic scene of the vehicle and help ensure a safe driving experience. To overcome the low accuracy issue encountered in past monocular depth estimation methods, a novel neural network model which combines conv_block and sep_block or identity_block and ap_block is proposed in this study. This model can ensure a sufficient depth of the neural network while reducing the computation cost and number of parameters by employing depthwise separable convolution. A combination of these approaches not only enables extraction of rich and accurate feature information by the neural network developed in this study but also allows fusion of low-level and advanced features with the help of skip-connection technique. Experimental tests were performed using the KITTI data set for validating the effectiveness of our proposed method. The experimental results show that the algorithm proposed in this study can improve the accuracy of monocular depth estimation.

Constructing neural networks in forms of block can simplify the overall structure and further improve the robustness of the model. We plan to make further adjustments based on the original block to improve the accuracy of the

network. The improvements in the neural network for depth estimation will benefit the perception of the traffic environment and assist the driving experience substantially. This is also the motivation for continuous advancement in monocular visual depth estimation.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant nos. 61502429 and 61505176), the Zhejiang Provincial Natural Science Foundation of China (Grant no. LY18F020012), the Zhejiang Open Foundation of the MOST Important Subjects, the China Postdoctoral Science Foundation (Grant no. 2015M581932), and the Zhejiang University of Science and Technology Graduate Research Innovation Fund (Grant no. 2019YJSKC04).

References

[1] F. Wang, C. Chao, and J. Huang, "A review of research on driverless vehicles," *China Water Transport Monthly*, vol. 16, no. 12, pp. 126–128, 2016.

[2] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1–3, pp. 7–42, 2002.

[3] J. Flynn, I. Neulander, J. Philbin et al., "Deepstereo: learning to predict new views from the world's imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5515–5524, Las Vegas, NV, USA, January 2016.

[4] A. Saxena, M. Min Sun, and A. Y. Ng, "Make3D: learning 3D scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, 2009.

[5] L. Ladicky, J. Shi, and M. Pollefeys, "Pulling things out of perspective," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 89–96, Columbus, OH, USA, June 2014.

[6] K. Karsch, C. Liu, and S. B. Kang, "Depth transfer: depth extraction from video using non-parametric sampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2144–2158, 2014.

[7] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international Conference on Computer Vision*, pp. 2650–2658, Tampa, FL, USA, December 2015.

[8] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5506–5514, Las Vegas, NV, USA, June 2016.

[9] I. Laina, C. Rupprecht, V. Belagiannis et al., "Deeper depth prediction with fully convolutional residual networks," in *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*, pp. 239–248, IEEE, Stanford, CA, USA, October 2016.

[10] P. Wang, X. Shen, Z. Lin et al., "Towards unified depth and semantic prediction from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2800–2809, Boston, MA, USA, June 2015.

[11] A. Chang, A. Dai, T. Funkhouser et al., "Matterport3D: learning from RGB-D data in indoor environments," 2017, http://arxiv.org/abs/1709.06158.

[12] S. Choi, D. Min, B. Ham, Y. Kim, C. Oh, and K. Sohn, "Depth analogy: data-driven approach for single image depth estimation using gradient samples," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5953–5966, 2015.

[13] J. Konrad, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Learning-based, automatic 2D-to-3D image and video conversion," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3485–3496, 2013.

[14] M. H. Baig and L. Torresani, "Coupled depth learning," in *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, Lake Placid, NY, USA, March 2016.

[15] J. Shi, X. Tao, L. Xu, and J. Jia, "Break ames room illusion: depth from general single images," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, p. 225, 2015.

[16] R. Ranftl, V. Vineet, Q. Chen et al., "Dense monocular depth estimation in complex dynamic scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4058–4066, Las Vegas, NV, USA, June 2016.

[17] R. Furukawa, R. Sagawa, H. Kawasaki et al., "Depth estimation using structured light flow—analysis of projected pattern flow on an object's surface," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4640–4648, Tampa, FL, USA, December 2017.

[18] C. Hane, L. Ladicky, and M. Pollefeys, "Direction matters: depth estimation with a surface normal classifier," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 381–389, Boston, MA, USA, June 2015.

[19] X. You, Q. Li, D. Tao et al., "Local metric learning for exemplar-based object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 8, pp. 1265–1276, 2014.

[20] W. Zhuo, M. Salzmann, X. He et al., "Indoor scene structure analysis for single image depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 614–622, Boston, MA, USA, June 2015.

[21] M. Liu, M. Salzmann, X. He et al., "Discrete-continuous depth estimation from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716–723, Columbus, OH, USA, June 2014.

[22] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[23] D. Eigen, C. Puhrsch, R. Fergus et al., "Depth map prediction from a single image using a multi-scale deep network," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 2366–2374, Montreal, Canada, December 2014.

[24] K. He, X. Zhang, S. Ren et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[25] M. Mancini, G. Costante, P. Valigi, T. A. Ciarfuglia, J. Delmerico, and D. Scaramuzza, "Toward domain independence for learning-based monocular depth estimation,"

IEEE Robotics and Automation Letters, vol. 2, no. 3, pp. 1778–1785, 2017.

[26] J. Xie, R. Girshick, A. Farhadi et al., "Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks," in *Proccedings of the European Conference on Computer Vision*, pp. 842–857, Springer, Amsterdam, Netherlands, October 2016.

[27] X. Yin, X. Wang, X. Du et al., "Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5870–5878, Tampa, FL, USA, December 2017.

[28] D. Xu, E. Ricci, W. Ouyang et al., "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5354–5362, Venice, Italy, October 2017.

[29] J. Li, R. Klein, A. Yao et al., "A two-streamed network for estimating fine-scaled depth maps from single RGB images," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3372–3380, Venice, Italy, October 2017.

[30] B. Li, Y. Dai, M. He et al., "Monocular depth estimation with hierarchical fusion of dilated CNNs and soft-weighted-sum inference," *Pattern Recognition*, vol. 83, pp. 328–339, 2018.

[31] A. Chakrabarti, J. Shao, G. Shakhnarovich et al., "Depth from a single image by harmonizing overcomplete local network predictions," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 2658–2666, Barcelona, Spain, December 2016.

[32] B. Ummenhofer, H. Zhou, J. Uhrig et al., "Demon: depth and motion network for learning monocular stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 5, pp. 5038–5047, Long Beach, CA, USA, July 2017.

[33] J. H. Lee, M. Heo, K. R. Kim et al., "Single-image depth estimation based on fourier domain analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 330–339, Long Beach, CA, USA, June 2018.

[34] H. Fu, M. Gong, C. Wang et al., "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2002–2011, Long Beach, CA, USA, June 2018.

[35] L. C. Chen, Y. Zhu, G. Papandreou et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–818, Munich, Germany, September 2018.

[36] Q. Zhou, W. Yang, G. Gao et al., "Multi-scale deep context convolutional neural networks for semantic segmentation," *World Wide Web*, vol. 22, no. 2, pp. 555–570, 2019.

[37] W. Yang, Q. Zhou, Y. Fan et al., "Deep context convolutional neural networks for semantic segmentation," in *Proceedings of the CCF Chinese Conference on Computer Vision*, pp. 696–704, Springer, Tianjin, China, October 2017.

[38] M. Menze, C. Heipke, and A. Geiger, "Joint 3D estimation of vehicles and scene flow," in *Proceedings of the ISPRS Workshop on Image Sequence Analysis (ISA)*, Taipei, Taiwan, August 2015.

[39] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014, http://arxiv.org/abs/1412.6980.

[40] M. Dimitrievski, B. Goossens, P. Veelaert et al., "High resolution depth reconstruction from monocular images and sparse point clouds using deep convolutional neural network," in *Proceedings of the Unconventional and Indirect Imaging, Image Reconstruction, and Wavefront Sensing 2017*, vol. 10410, p. 104100H, International Society for Optics and Photonics, San Diego, CA, USA, August 2017.