OptiDepthNet : A Real-time Unsupervised Monocular Depth Estimation Network

Feng Wei (lygweifeng@126.com)
Hohai University - Jiangning Campus https://orcid.org/0000-0001-5540-4449
XingHui Yin
Hohai University - Jiangning Campus
Jie Shen
Hohai University - Jiangning Campus
HuiBin Wang
Hohai University - Jiangning Campus

Research Article

Keywords: OptiDepthNet,Monocular depth estimation,Depth separable convolution,Kitti,Depth learning

Posted Date: December 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-812743/v1

OptiDepthNet:A Real-time Unsupervised Monocular Depth Estimation Network

Feng Wei[1] · XingHui Yin[2] · Jie Shen[3] · HuBin Wang[4]✉

Abstract

With the development of depth learning, the accuracy and effect of the algorithm applied to monocular depth estimation have been greatly improved, but the existing algorithms need a lot of computing resources. At present, how to apply the existing algorithms to UAV and its small robot is an urgent need.Based on full convolution neural network and Kitti dataset, this paper uses deep separable convolution to optimize the network architecture, reduce training parameters and improve computing speed. Experimental results show that our method is very effective and has a certain reference value in the development direction of monocular depth estimation algorithm.

Keywords OptiDepthNet·Monocular depth estimation·Depth separable convolution·Kitti·Depth learning·

1 Introduction

With the development of artificial intelligence technology, computer vision technology is widely used in industrial sites, such as production line inspection, UAV detection, cooperative robot and intelligent driving.A good understanding of scene information helps the robot to accurately locate and complete complex technical actions. Accurate and effective depth information can improve the effects of 3D reconstruction, target recognition and semantic segmentation [1].

At present, there are many ways to obtain depth information, which can be divided into active and passive methods. Active methods mainly use ultrasonic, laser TOF, lidar and so on. They rely on the sensor to send signals to obtain the depth information of objects in the scene. The depth information is obtained quickly, but there are some disadvantages, such as large sensor volume, high power consumption, and the measured data are easy to be disturbed by external noise and other environmental interference. Passive methods are common, such as binocular stereo matching [2] and motion recovery structure [3] multiview stereo matching, which saves cost, but requires camera parameter calibration and a large number of algorithm calculations, has certain requirements for the hardware platform, takes a certain amount of time,

✉HuiBin Wang,
hbwang@hhu.edu.cn
[1] School of computer and information, Hohai University,Nanjing 211100,China
[2] School of computer and information, Hohai University,Nanjing 211100,China
[3] School of computer and information, Hohai University,Nanjing 211100,China
[4] School of computer and information, Hohai University,Nanjing 211100,China

and the calculation accuracy is greatly affected by the environment. At present, these common depth acquisition methods have limitations for some small robot platforms, and have certain limitations in power consumption, volume and cost. However, the scheme based on single purpose depth estimation has become a choice. Camera imaging is to convert three-dimensional information into two-dimensional information, and the depth information is lost in the imaging process. If only a single image is used to restore the depth information of the image, it has been regarded as an ill conditioned problem and is difficult to achieve.

At present, the depth learning method has been gradually applied to the field of monocular depth estimation. The commonly used methods mainly include unsupervised learning method and supervised learning method.The supervised learning method is to build a codec network, extract depth information, input a pixel level image and output the corresponding depth map and label, and transform the depth estimation problem into a pixel level regression problem. However, the supervised learning method relies on large sample data set training, which is expensive and has great training pressure. It can not achieve good classification and recognition for some unfamiliar scenes, which limits the popularization and use of the supervised learning method. Unsupervised learning methods can effectively avoid the training pressure of big data and make up for these defects. However, in order to improve the depth estimation accuracy, many unsupervised learning methods are committed to increasing the depth and feature extraction details of the deep neural network architecture, resulting in an increase in the number of network parameters, which requires the platform to have higher computing power and storage capacity. With the deepening of research, how to apply these excellent algorithms and architectures to embedded devices with limited hardware resources has become an important research direction. One of the important challenges is how to maintain the balance between the accuracy of monocular depth estimation algorithm and improve the computing speed.

At present, the common method is to use the embedded platform to obtain images, transmit them to the edge server for training after preprocessing, and then transmit the trained data and models to the embedded platform for depth estimation [4]. In this method, the embedded device terminal only performs the functions of acquisition and communication, and does not really implement the algorithm on the end side.To solve this problem, this paper proposes an unsupervised learning method based on optimization, which can effectively reduce the number of network parameters and floating-point operations under the condition of ensuring that the output depth information remains unchanged, help to improve the computing power of the platform, realize the termination of the depth learning algorithm, and contribute to the distributed computing of the overall architecture.

The main idea of this algorithm is to introduce deep separable convolution into the codec architecture, improve the computing power of the network by optimizing the parameters of the convolution layer, and realize feature extraction and image reconstruction. Practice has proved that after running the optimized self coding network on NVIDIA geforce RTX 2080 super platform, the training speed is increased by more than 33.3%, and the image accuracy remains at the original level and is slightly improved. Specifically, our contributions are as follows:

1) A full convolution unsupervised monocular depth estimation model OptDepthNet is proposed to perform left-right depth consistency. The encoder is based on resnet50 architecture model and uses depth separable convolution instead of ordinary convolution for optimization;

2) The performance comparison with several common models shows the effectiveness of our

method, reduces network parameters and improves operation efficiency.

To sum up, our contribution is to propose an optimized monocular depth estimation method, which optimizes the architecture of depth neural network to realize that the algorithm can run on embedded platform.

2 Related work

In the research of monocular depth estimation, depth learning is the most advanced method of depth estimation based on RGB images, which is trained with large-scale data sets such as Kitti. In training, supervised learning method requires each RGB image to have its corresponding depth label, while unsupervised method does not need depth label. Its basic idea is to solve the relative depth from the object to the camera in the image by using the left and right views and the idea of epipolar geometry.

According to the research methods, we understand the development of supervised and unsupervised methods, and introduce the development of real-time network architecture suitable for embedded devices.

2.1 Supervised Depth Estimation

For the depth estimation task of a single image, in many cases, we pay attention to the prediction of absolute depth. In particular, the working sites such as industrial robots and UAVs need to select the working strategy according to the scene depth. Generally, the supervised regression model is used for prediction, that is, the model training data is labeled, and the continuous depth values can be regressed and fitted.

Eigen et al. (2014) first introduced depth learning into the field of monocular depth estimation and proposed coarse-scale network and fine-scale network architecture. The coarse-scale network is used to predict the global depth of the scene and obtain depth clues such as target location, vanishing point and spatial alignment, and the fine-scale network is used to locally optimize the results of global prediction [5].Based on this research, eigen et al. (2015) proposed a unified multi-scale network architecture, using a deeper VggNet network, using three fine scales to increase details and improve resolution, adding gradient regularization term on the basis of scale invariant loss, and calculating the difference between predicted gradient and real gradient for depth prediction Surface normal vector estimation and semantic segmentation [6].Liu et al. (2015) combined the depth convolution network with the continuous random field, used the univariate potential energy and paired potential energy term of the continuous CRF, and used the depth structured strategy to extract the estimated depth [7]. Li et al. (2015) proposed a multi-scale depth estimation method. Firstly, the super-pixel scale is regressed by neural network, and then the multi-layer conditional random field post-processing is used to optimize the depth of super-pixel scale and pixel scale [8]. Laina et al. (2016) added residual learning to the full convolution network architecture, increased the depth of the network structure to improve the depth estimation effect, and proposed a new up-sampling method and BerHu as the loss function [9]. Cao et al. (2018) treated the depth estimation problem as a pixel level classification problem, projected the depth value into the logarithmic space, and then discretized the continuous depth

value into category labels according to the depth range [10].

Although supervised depth estimation can obtain better depth estimation accuracy, each image is required to have the corresponding label depth, and the acquisition price of depth label is very expensive, and the collected original depth label is usually some sparse points, which can not match the original image well.

2.2 Unsupervised Depth Estimation

The unsupervised method does not need depth labels. The existing left and right view sets can meet the research requirements, and the relative depth from the object to the camera can be obtained by combining the polar constraints and the automatic coding mechanism.

Garg et al. (2016) used the original image and the target image to form a stereo image pair. First, the encoder was used to predict the depth map of the original image, and then the decoder was used to reconstruct the original image combined with the target image and the predicted depth map, and the reconstructed image was compared with the original image to calculate the loss [11]. Godard C et al. [2017] realized unsupervised depth prediction by using the consistency of left and right views, generated parallax map by using epipolar geometric constraints, improved performance and robustness by using the consistency of left and right views, learned the mapping relationship from left (right) image to right (left) image, estimated the scene depth information, and transformed monocular image depth estimation into image reconstruction [12].Godard C et al. Subsequently added the loss of left and right image consistency and the loss of enhanced parallax smoothness on the basis of this research, which further improved the upgraded network effect and the accuracy of depth information estimation, but still did not solve the problems such as unclear object contour and unsmooth depth change in the obtained depth map [13]. Tosi et al. [14] transformed monocular depth estimation into stereo matching problem, and then used stereo matching network for parallax estimation. The whole network structure includes the following parts: primary feature extraction network, primary parallax estimation network and parallax optimization network.Casser et al. [15] proposed that by modeling the scene and a single object, introducing geometry in the learning process, and self-learning the camera's self motion and object motion. Wang et al. [16] proposed an idea of calculating the loss function in the hierarchical embedding space for depth estimation model training. On the one hand, a generator HEGS for generating multi-level embedding is designed to extract features from the depth map and construct subspaces at different levels. Then, the loss function is constructed by calculating the distance between the reference depth embedding and the predicted depth embedding.Mancini et al. [17] proposed a visual object detection system, which uses the depth neural network method to train real images and synthetic images to realize depth estimation, and can detect obstacles at long distance and high speed. Amir et al. [18] proposed a training method based on style conversion and antagonistic training. Based on the training of a large number of synthetic environment data, the depth of pixels is predicted from a single real color, but this method can not be applied to the sudden illumination change and saturation in style conversion.

These unsupervised methods basically obtain higher depth map accuracy by increasing the complexity of the network. The network parameters are too large, and the calculation needs a lot of resources.

2.3 Lightweight Monocular Depth Estimation Network

With the increase of network complexity and computation of unsupervised methods, although the accuracy of depth map is getting better and better, these algorithms can not be applied to small robot platforms with limited resources. At present, there is an urgent need to optimize the complexity of existing algorithms, reduce training parameters and take into account the accuracy of image acquisition. At present, the main idea is to improve and optimize the network structure.

Fast target detection and classification methods in deep learning are conducive to image semantic segmentation. Common detection models include SSD [19], Yolo3 [20], and classification schemes also include AleNet [22], Vgg [23], ResNet [24]. SSD combines the advantages of Yolo and FastRCNN [21], with fast speed and high accuracy.

Dianna wofk et al. (2019) for embedded system equipment, mobilenet2 [25] is used in the encoder part, and deep separable network is also introduced in the decoder part. Its original intention is to lighten the codec structure, and network pruning [26] and other technologies are adopted to reduce training parameters and memory usage [27]. Finally, the execution force is 27fps on Jetson TX2 CPU, and the parameters are 1.34m, The accuracy of Vgg is similar.Jun Liu et al. (2020) proposed a MiniNet network structure with recursive function [28], which not only maintains the extremely light size, but also realizes the ability of a deep network, but also maintains the real-time high-performance unsupervised single-sided depth prediction of video sequences, and can realize the rate of 110 frames per second on a single GPU, 37 frames per second on a single CPU and 2 frames per second on raspberry PI3.

In this paper, we propose a lightweight network OptiDepthNet, which is based on the existing full convolutional codec network and introduces the deep separable network optimization technology, which greatly improves the running speed of training while ensuring the accuracy effect.

3 Method

This section introduces our network architecture of unsupervised single image depth estimation. Inspired by the U-Net network [29] and DeeperLab3 structure [30], we introduce a layer hopping structure between the encoder and decoder, and introduce depth separable convolution into the encoder and decoder to improve the network computing speed, realize the balance between estimation accuracy and computing speed, and learn from the principle of image depth acquisition We analyze our optimized network from the aspects of codec network, reconstruction of depth image and so on.

3.1 Obtaining Depth Estimation From Image Reconstruction

In the test case, inputting an image I_{in} can output the corresponding depth map d_{out}, which requires us to realize $d_{out} = f(I_{in})$ according to a calculation function F. Usually, in the

process of obtaining function F , we construct an unsupervised learning scheme according to the principle of binocular ranging, and realize depth image reconstruction by combining training loss and left-right consistency check.Assuming that the image is corrected, the baseline distance between the two cameras is b, the camera focal length is F , and Δd is the image parallax of the left and right input images I_l and I_R . According to the parallax acquisition formula $d_{out} = b * f / \Delta d$, the depth d_{out} of the pixel can be preliminarily obtained [31]. According to the full convolution network architecture [32], the calibrated image pairs are input into the training network, combined with left-right consistency loss, parallax smoothing loss and appearance matching loss, so as to realize network training and obtain a good model architecture.

3.2 Network architecture

We propose that the OptiDepthNet network belongs to the full convolution network architecture, and refer to the left-right consistency network proposed by Godard et al. However, there are some modifications in the network architecture, so that we can greatly improve the speed in the training network. The network structure is shown in Figure 1. Our network is mainly composed of an encoder and decoder, which realizes depth map reconstruction and semantic segmentation for the input image. The features extracted from different layers of the encoder are fused in the decoder to improve the detail and feature accuracy of the reconstructed depth map. The parallax map is generated according to the left and right images, and the image reconstruction is realized through the depth neural network. The output depth image does not represent the absolute distance from the object in the image to the camera, but the relationship between the objects in the image. The brighter the brightness in the figure, the closer it is to the camera.

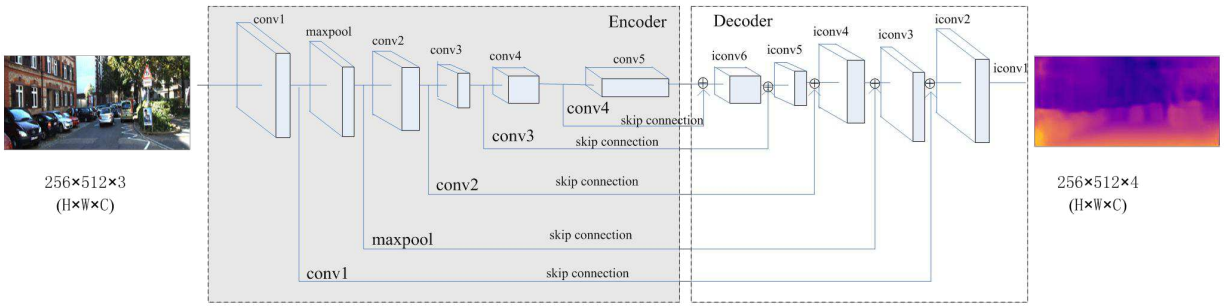


Figure 1: RGB image input coding layer extracts features and enters the decoder for image reconstruction to generate depth map.

The encoder is responsible for extracting the depth features of the input image, while the decoder gradually restores the details and corresponding spatial dimensions of the target through up-sampling and deconvolution, and uses skip-connection to compensate for the loss of some features, so as to achieve the reconstruction of the depth image. Because our basic purpose is to obtain real-time depth estimation, and extracting rich image features is very important for accurate depth prediction, we choose the classical residual network structure

Resnet50 as the main framework of the encoder, and introduce depth separable convolution on this basis.

3.2.1 Encoder Network

In recent years, with the deepening of CNN network to solve more complex practical problems, it is also accompanied by some gradient disappearance and gradient explosion, which makes training very difficult. Our coder DResnet is optimized based on Resnet50 and consists of a standard convolution layer and four groups of residual blocks.The first layer of the encoder is a 7 ×7 convolution in steps of 2, then activated by the ELU function, and the number of output channels is set to 64. In the residual block, the middle convolution part is changed to depth separable convolution, and the convolution size of the rest is 1.

Fig. 2 (a) shows a normal residual block model, for the input image I_x , three convolution operations are performed:64 convolutions of 1×1 and 3×3, and 256 convolutions of 1×1 , extract the feature to obtain the output feature I_{x1} ,the input I_x is connected to the output through shortcoming, and the output I_y through ELU. Fig. 2 (b) shows the convolution module with a convolution layer of 3×3 is optimized into a deep separable convolution, and the I_{x1} part is divided into three groups according to the number of channel: then the convolution of 3×3 ,the convolution of 1×1 , and finally output the characteristic graph I_{x2} .

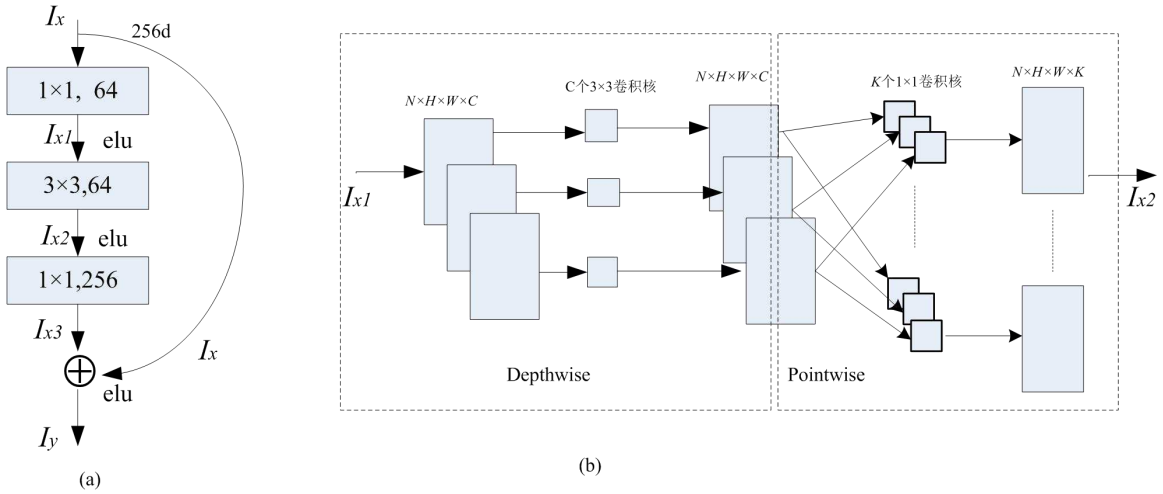


Fig. 2: (a) implementation process of residual block convolution, (b) implementation process of changing middle part convolution to deep separable convolution.

The normal convolutional of 3×3 operates with the convolution of 3×3 and 1×1 respectively, and deep separable convolution achieves channel and region separation, with great improvement in computational performance, reducing training parameters, but the channel has a similar output effect.

3.2.2 Decoder Network

The function of the decoder is to reconstruct the extracted feature map of the encoder to form dense prediction and obtain the depth map corresponding to the input. Each layer of the encoder is used to gradually reduce the spatial resolution and extract higher-level features. Many image details may be lost, which makes it difficult for the decoder to recover pixel level data.In order to meet the requirements of high precision and real-time, the deep separable convolution operation is performed in the output part of each stage to simplify the network parameters, as shown in Fig. 3.The output of the encoder is regarded as the input of the first layer of the decoder, after the nearest neighbor interpolation method and the up-sampling with scale 2,upconv6 after convolution of 3×3,also fuse the conv4 layer of the encoder as output. Then through the deeply separable convolution of depthwise and pointwise,the computational parameters are greatly reduced in the output layer to achieve network lightweight, and finally obtain the input of the decoder of the next level of decoder.

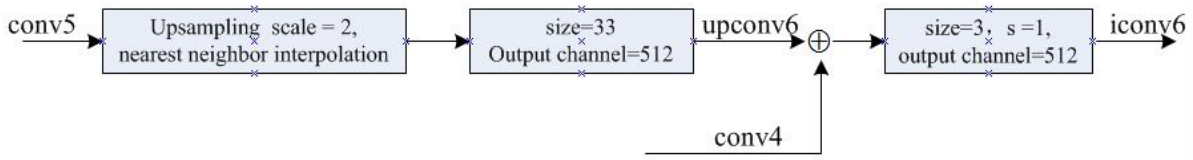


Fig. 3: The output signal of conv5 is upsampled and obtain upconv6 after the the convolution of 3×3,connect conv4 as output, and then conduct depthwise and pointwise convolution to obtain the output of iconv6 as the input of the next decoder.

Our decoder of DDensenet consists of five fusion modules and reduces the number of output channels by half relative to the number of input channels.Through the methods of interpolation and deconvolution, the feature maps with the size of the original image、1/2、1/4、1/8、1/16 and 1/32 are obtained. Then the feature maps of these six sizes are concatenated with the feature maps of the same size obtained in the original encoder to generate six size parallax maps.

3.3 Loss Function

The algorithm in this paper follows the loss function proposed in reference [12], which is composed of three parts: the similarity between the generated reconstructed image and the original image, the smoothness of the parallax image and the consistency of the predicted left and right images.The appearance matching loss indicates that the input left and right parallax images need to be sampled by parallax in the training network, and then the images are generated by bilinear sampling, which is composed of *L1* regularization and *SSIM* [33].The function of parallax smoothing loss is that parallax becomes smooth [34], and the generated parallax map can be as continuous as possible through *L1* regularization.The loss of left and right consistency is the loss of the consistency of the left and right disparity map. When only the left view is input, the left and right disparity map is predicted. In order to ensure the consistency, the

consistency penalty of *L1* left and right disparity map is used as a part of the model.

4 Experiment

In this section, we will use the Kitti data set experimental results to prove the effectiveness of our method, and carry out various model measurements on the existing research. Comparing various encoders, according to the execution efficiency and image data, we also show that deepening the depth of the network structure can improve the image quality to a certain extent.

4.1 Kitti Dataset

Kitti Dataset is used to evaluate the performance of computer vision technologies such as stereo image, optical flow and visual ranging in vehicle environment, including real image data collected from urban, rural and expressway scenes. 3756 frames are selected from 30 scenes for training and 500 frames are used for verification.In this paper, the image resolution of each input RGB frame is adjusted to 256×512 pixels, 256×512 pixels for depth map output.

4.2 Implementation Rules

The depth estimation network of OptiDepthNet proposed by us is implemented on the open tensorflow model. Our network is trained on the Kitti dataset, and their accuracy is evaluated by official training and test data segmentation.For training, we use one GPU with 23500 training steps and 256×512 image pixels , the batchsize is 8, and the initial learning speed is learning_ rate is set to 0.0001, num_ threads adopts 8.Our OptiDepthNet is trained on an i7-9700 CPU, the main frequency is 3G, the RAM is 32G, and the graphics card is NVIDIA geforce RTX 2080 super,then the whole training time is 36 hours.In the training process, the input frame is uniformly sampled with a probability of [0.8,1,2] for color and saturation ,and [0.5,2.0] for brightness with a probability of 50% to implement image enhancement.

In the process of building the network model, we use DResnet50, a variant of Resnet50 model, as our encoder, and the architecture and training process of other models remain unchanged.

Based on previous work, we used several image evaluation indexes to evaluate the depth images obtained by our OptiDepthNet in unsupervised monocular depth estimation [35].The quantitative evaluation indexes in monocular image depth estimation are relative error(REL), root mean square error(RMS), log error(LG), and accuracy(% correct) used by most algorithms. Generally, the smaller the error, the better, and the higher the accuracy, the better.

4.3 Experimental Results

Firstly, we compare the related work of this part qualitatively and quantitatively. Secondly, we analyze the computational efficiency of our OptiDepthNet. Thirdly, we give the effect of depth

estimation on Kitti dataset. Fourthly, we study the deepening of network level to prove the effectiveness of our optimization method.

4.3.1 Comparison With Other Work

We evaluated OptiDepthNet using Kitti split, and the results of our method are listed in Table 1. At the same time, compared with Resnet50 as encoder and Densenet as decoder [36], our method is to optimize the convolution operation of encoder and decoder to reduce the amount of calculation, and the effect is obvious.

Mothod	Setting	Error (lower is better)			Accuracy (higher is better)			Parameters
	cap	Abs Rel	Sq Rel	RMSE	δ1	δ2	δ3	
Kuznietsov et al. [37]	0-80m	0.308	9.367	8.700	0.752	0.904	0.952	80.84 M
Zhou et al. [38]	0-80m	0.208	1.768	6.856	0.678	0.885	0.957	34.20 M
Yin et al. [39]	0-80m	0.155	1.296	5.857	0.793	0.931	0.973	58.45 M
Gordon et al. [12] +Resnet50	0-80m	0.1495	1.5606	6.851	0.783	0.900	0.950	58.4M
Ours+resnet50	0-80m	0.1417	1.3602	6.339	0.792	0.916	0.963	30.36M
Gordon et al. [12] +Vgg	0-80m	0.1843	2.1966	8.230	0.721	0.854	0.924	31.6M
0urs+Vgg	0-80m	0.1814	2.0958	7.563	0.745	0.880	0.944	3.58M

Table 1: The input 256×512 pixel images are compared with various codec networks.

As can be seen from table 1, our OptiDepthNet network parameters are reduced by 2.67 times compared with kuznietsov. Gordon [12], which is closely related to us, is selected for detailed comparison. As shown in Figure 4, our network parameters are 1.9 times less than Gordon's network. If the encoder selects Vgg network, the parameters used in our method are 8.28 times less than Gordon's method, and 1.54 times less than those using Resnet152, It can be seen that if the encoder adopts a lightweight network, more parameters will be reduced. The comparative evaluation index is that the image quality obtained by using Resnet50 as encoder is better.

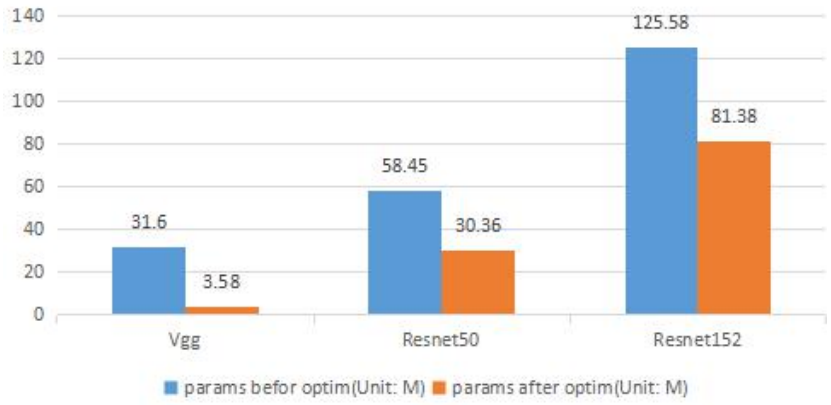


Figure 4: Input image Comparing the parameters before and after network optimization, the encoder adopts Vgg,Resnet50 an Resnet152 networks respectively. After the optimization, the parameters of the network with fewer parameters are reduced more after optimization.

The running model is trained on a single GPU and compared with Gordon's work. The running time of a single epoch is tested with Vgg, Reset50 and Resnet152 as encoders respectively. DVgg, DResnet50 and DResnet152 are optimized networks respectively. Figure 5 summarizes the final results obtained by our method and the comparison results with previous work.

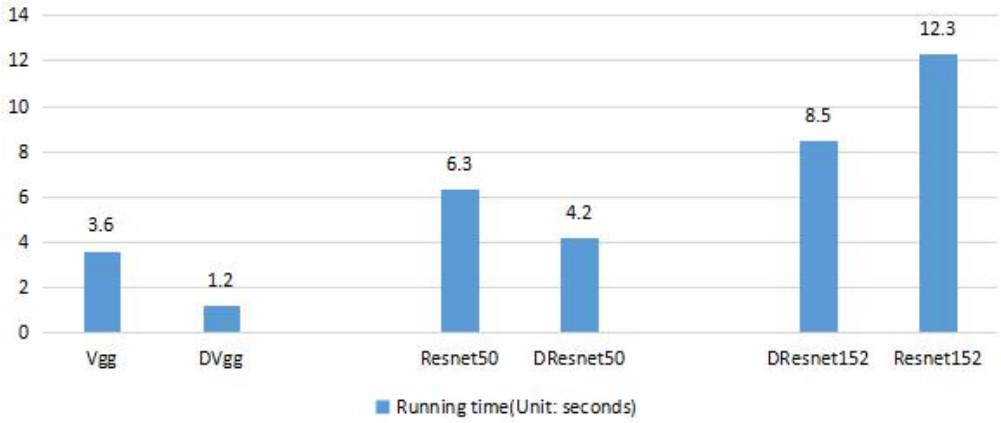
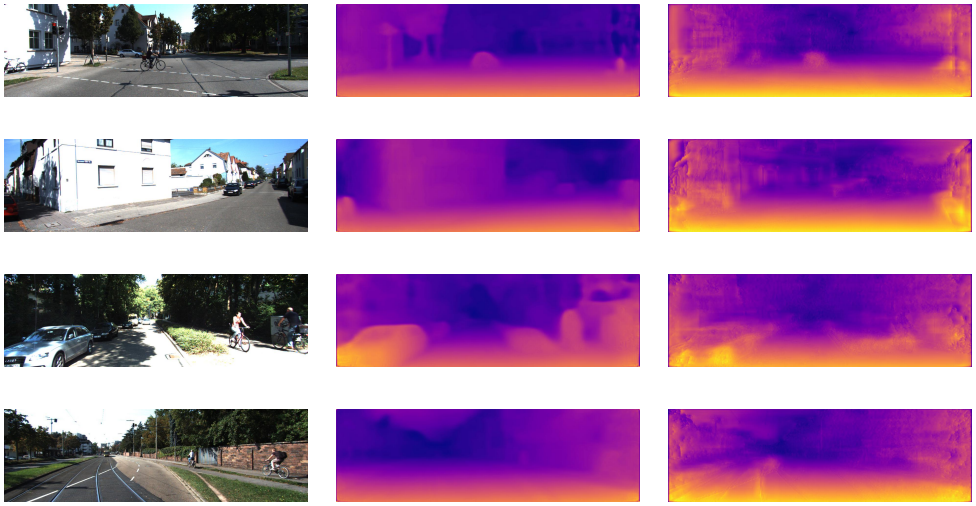


Figure 5. Our method saves a lot of time in feature extraction and depth map reconstruction. The time of per epoch is listed at the top of the histogram, which is twice as long as the network running time of DVgg architecture.

From the above comparison data, it can be seen that using DResnet50 as encoder and DDensetnet as decoder for image reconstruction has made a great leap in network parameters and computing speed, which is conducive to the embeddedness of the network.

4.3.2 Kitti Dataset Test

Figure 6 compares and analyzes the quality of four images in Kitti. The encoder analyzes Resnet50 and DResnet50 respectively. It can be seen that the optimized image depth is prominent. Table 2 shows the comparison of image quality results of four images before and after optimization.



(a) (b) (c)

Fig. 6: Monocular depth estimation results based on Kitti dataset, (a) the original figure, (b) the encoder is the depth map of Resnet50, (c) the encoder is the depth map of DResnet50 network optimized by Resnet50.

encoder	Abs	Sql	RMS	Abs	δ1
Resnet50	0.1495	1.5606	6.851	0.1495	0.783
DResnet50	0.1417	1.3602	6.339	0.1417	0.792

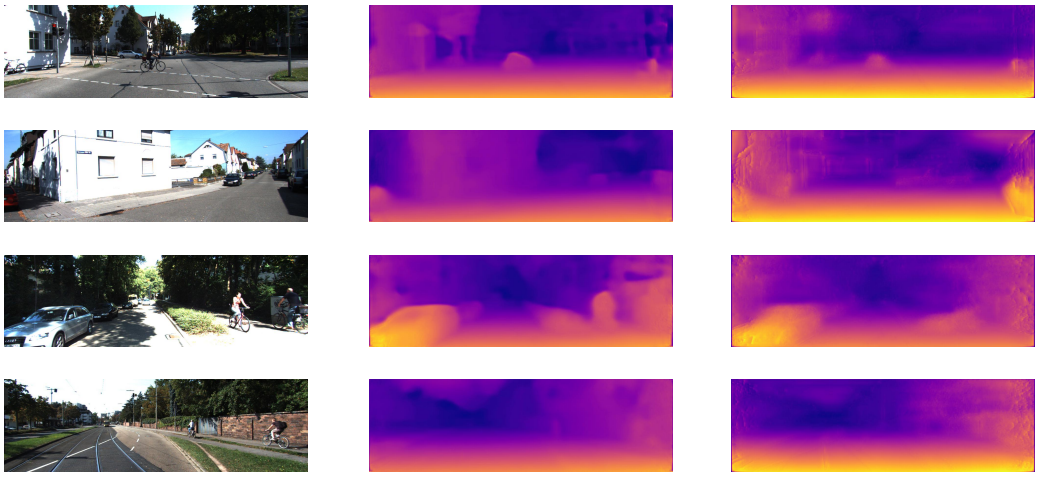
Table 2: Result quality comparison of four test images in Kitti with encoders Resnt50 and DResnet50 respectively.

From the test samples taken immediately, it can be seen that the optimized network model has a certain optimization in image quality.

4.3.3 Extension To Othe Network Structures

In Figure 7, we show the extension of the network optimization method to the network model. Compared with Vgg for the encoder and DVgg for the deep separable convolution network optimization network structure, the network parameters and single operation time have been greatly improved. For the generated model, four images in Kitti dataset are used to test respectively. The image quality is shown in Table 3, and the image quality has been improved to a

certain extent.



(a) (b) (c)

Fig. 7: Monocular depth estimation results based on Kitti dataset, (a)the original figure, (b) the depth figure with Vgg encoder, (c)the depth figure with Vgg optimizedDVgg network encoder.

Encoder	Abs	Sql	RMS	Abs	δ1
Vgg	0.1843	2.1966	8.230	0.1843	0.721
DVgg	0.1814	2.0958	7.563	0.1814	0.745

Table 3:The encoder is trained and tested by Vgg and DVgg networks respectively, and the training quality of DVgg network is improved to a certain extent.

It can be seen from the comparison that after Vgg network optimization, the image performance parameters are also improved and optimized. It can be concluded that our optimization method is also applicable to a variety of network architectures.

4.4 Limitations

This method has made great progress in improving network parameters and running speed, but there are also some problems. There is a great improvement in small object extraction and contour, but the depth image leads to the loss of some details of the reconstructed boundary, so the reconstruction algorithm needs to be further improved. Network optimization technology needs to be further improved in memory usage. Network pruning can be used to reduce memory consumption. It is also necessary to further optimize and enhance the image edges and details. Good image quality can enable the robot to quickly identify and classify, and ensure the normal operation of small embedded devices such as robots and UAVs.

5 Results

The monocular depth estimation method of unsupervised depth neural network proposed by us is tested and tested on the binocular correction data set based on Kitti. The depth image obtained is not the absolute distance from the object to the camera, but only represents the relative distance from various objects in the image to the camera. Later, we hope to make a breakthrough in measuring absolute distance with our method.

The method of this paper focuses on the improvement of network parameters and network training speed, which increases the real-time performance of the network, so that it can carry out real-time observation in embedded devices, including robots, UAVs and other small devices. At present, the main work is to test a single image. We hope that our optimization method can adapt to video processing, and our method can make the deep learning method perform well in embedded devices.

Thank Godard and others for sharing their team's research results and codes, as well as David Eigen and others for their papers and research results.

Funding. This work has been completed in the scope of NSFC, it is supported by the National Natural Science Foundation of China According to No. 61903124.

Declarations

Confict of interest The authors declare no confict of interest.

References：

1. Liu F Y,Shen C H,Lin G S,Lin G S and Reid I.2016.Learning depth from single monocular images using deep convolutional neural fields.IEEE Transactions on Pattern Analysis and Machine Intelli-gence，38 (10) : 2024-2039 ［DOI: 10． 1109 / TPAMI． 2015．2505283］
2. Zhu Qingbo, Wang Hongyuan.Block recovery stereo matching algorithm using image segmentation [J].Journal of Huazhong University of Science and Technology (Natural Science Edition),2010,38(1)：81-84．
3. Xie Zexiao, Zhou Zuoqi.Spatial point localization method based on the Motion Recovery Structure [J]. Progress in Laser and optoelectronics,2018,55 (8): 370-377.
4. Xu Cheng, Tu Xiaohan, Liu Siping, et al.Fast monocular depth estimation methods for embedded platforms. CN110599533A[P].2019.
5. Eigen.D, C. Puhrsch, and R. Fergus.Depth map prediction from a single image using a multi-scale deep network.in Advances in Neural Information Processing Systems (NIPS),

2014, pp. 2366–2374.
6. Eigen D,Fergus R.Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture.2015 IEEE International Conference on Computer Vision (ICCV), 2014.
7. F. Liu,C. Shen, and G.Lin. Deep convolutional neural fifields for depth estimation from a single image. in Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5162–5170.
8. Li B, Shen C H, Dai Y C, Van den H A and He M Y.2015.Depth andsurface normal estimation from monocular images using regression on deep features and hierarchical CRFs. Proceedings of 2015 IEEEConference on Computer Vision and Pattern Recognition. Boston:IEEE,1119-1127［DOI: 10．1109 / CVPR．2015．7298715］
9. I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks.in International Conference on 3D Vision (3DV), 2016, pp. 239–248.
10. Cao Y , Wu Z , Shen C.Estimating Depth From Monocular Images as Classification Using Deep Fully Convolutional Residual Networks[J].IEEE Transactions on Circuits and Systems for Video Technology, 2018.
11. Garg R, Vijay Kumar B G, Carneiro G and Ian R.2016.Unsupervised CNN for single view depth estimation:geometry to the rescue.Pro-ceedings of the 14th European Conference on Computer Vision.Am-sterdam: Springer,740-756DOI: 10．1007 /978-3-319-46484-8 45］
12. C. Godard,O. Mac Aodha,and G. J. Brostow.Unsupervised monocular depth estimation with left-right consistency.in Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
13. Godard C,Aodha O M,Firman M,et al.Digging Into Self-Supervised Monocular Depth Estimation[J].2018.
14. F.Tosi,F.Aleotti,M. Poggi, S.Mattoccia.Learning monocular depth estimation infusing traditional stereo knowledge.in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9799–9809. doi:10.1109/CVPR.2019.01003
15. Casser V,Pirk S,Mahjourian R,et al.Depth Prediction Without the Sensors:Leveraging Structure for Unsupervised Learning from Monocular Videos[J]. 2018.
16. Wang L , Zhang J , Wang Y , et al. CLIFFNet for Monocular Depth Estimation with Hierarchical Embedding Loss[M]. Springer, Cham, 2020.
17. Mancini M,Costante G,Valigi P,et al.Fast Robust Monocular Depth Estimation for Obstacle Detection with Fully Convolutional Networks[J]. IEEE, 2016.
18. Atapour-Ab Arghouei A.Real-time monocular depth estimation using synthetic data with domain adaptation.[C].IEEE/CVF Conference on Computer Vision & Pattern Recognition. IEEE, 2018.
19. Liu W,Anguelov D,Erhan D,et al. SSD: Single Shot MultiBox Detector[J]. European Conference on Computer Vision, 2016.
20. Redmon J , Farhadi A . YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.
21. Girshick R . Fast R-CNN[J]. arXiv e-prints, 2015.
22. Technicolor T,Related S,Technicolor T,et al.ImageNet Classification with Deep Convolutional Neural Networks [50].
23. Lecun Y.Optimal Brain Damage[J].Neural Information Proceeding Systems,1990, 2(279):

598-605.
24. He K,Zhang X, Ren S, et al.Deep Residual Learning for Image Recognition[J].2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
25. M. Mancini,G.Costante,P. Valigi, and T. A. Ciarfuglia. Fast robust monocular depth estimation for obstacle detection with fully convolutional networks.in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016, pp. 4296–4303.
26. Lecun Y.Optimal Brain Damage[J].Neural Information Proceeding Systems.1990, 2(279): 598-605.
27. Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J]. 2017.
28. Jlab C , Qlab C , Rui C , et al. MiniNet: An extremely lightweight convolutional neural network for real-time unsupervised monocular depth estimation[J].ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 166:255-267.
29. Ronneberger O,Fischer P,Brox T.U-Net:Convolutional Networks for Biomedical Image Segmentation[J]. Springer International Publishing, 2015.
30. Chen L C,Zhu Y , Papandreou G,et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation[C]. European Conference on Computer Vision. Springer, Cham, 2018.
31. Cui Enkun,Teng Yanqing,Liu Jiawei.Calibration error compensation for the stereo measurement system [J]. Applied Optics, 2020, v.41;No.242(06):46-52.
32. Johnson J , Karpathy A , Li F F .Fully Convolutional Networks for Semantic Segmentation.
33. Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli.Image quality assess ment: from error visibility to structural similarity. IEEE transactions on image processing 13 (4) (2004) 600–612. doi:10.1109/TIP.2003.819861.
34. P. Heise, S. Klose, B. Jensen, and A. Knoll. Pm-huber: Patchmatch with huber regularization for stereo matching. In ICCV, 2013. 4
35. A. Ranjan, V.Jampani,L. Balles, K.Kim, D. Sun, J. Wulffff, M. J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12240–12249. doi:10.1109/CVPR.2019.01252
36. Diana Wofk , Fangchang Ma , Tien-Ju Yang, Sertac Karaman, and Vivienne Sze .FastDepth: Fast Monocular Depth Estimation on Embedded Systems.International Conference on Robotics and Automation (ICRA),2019
37. Y. Kuznietsov, J. Stuckler, B. Leibe.Semi-supervised deep learning for monocular depth map prediction. in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6647–6655. doi: 10.1109/CVPR.2017.238.
38. T. Zhou, M. Brown, N. Snavely, D. G. Lowe.Unsupervised learning of depth and ego-motion from video. in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, pp. 6612–6619. doi:10.1109/CVPR. 2017.700.
39. Z. Yin, J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, 2018, pp. 1983–1992. doi: 10.1109/CVPR.2018.00212.