https://doi.org/10.1016/j.asoc.2020.106272

Semi-supervised image depth prediction with deep learning and binocular algorithms

Kuo-Kun Tseng [a]
[a] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen, Shenzhen, China

Yaqi Zhang [a]
[a] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen, Shenzhen, China

Qinglin Zhu [a]
[a] School of Computer Science and Technology, Harbin Institute of Technology Shenzhen, Shenzhen, China

K.L. Yung [b]
[b] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, China

W.H. Ip [b]
[b] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, China

Abstract

Combining the advantages and disadvantages of supervised learning and unsupervised learning strategies in convolution neural networks, this paper proposes a semi-supervised single-image depth prediction model based on binocular information and sparse laser data. The model improves the depth prediction accuracy by introducing sparse depth monitoring information, which provides a better convergence of the model with a local optimal solution. In the experiment, we validate the effectiveness of the model on the KITTI data set. Compared to the supervised algorithm, the root mean square error is reduced by 41.6% and, compared to the unsupervised algorithm, the root mean square error is reduced by 26.9%.

1. Introduction

Image depth information plays an important role in object recognition, scene restoration, autonomous driving and other fields. With the increasing demand for new technologies, the rapid development of convolution neural networks (CNNs) in recent years has provided a new breakthrough in the development of visual depth technology. The existing research results can be divided into supervised and unsupervised methods.

Among the existing research methods and supervised methods, the most classic model is the network model proposed by the Eigen Group from New York University in 2014 [1], which divides the depth prediction into two steps. The first step is to predict the overall coarse-grained depth map. The second step is to fine-tune the resulting coarse-grained depth information result map, corresponding to a coarse network and a refined network. At CVPR 2015, Laina et al. [2] proposed a depth prediction model based on the Fully Convolutional Network (FCN) [3] structure. The existing depth prediction with a deep learning-based approach generally uses an end-to-end framework. That is, it employs convolutional upsampling for feature extraction and analysis. The parameters are trained to reduce the error between the prediction and the true value.

The core of an unsupervised algorithm is the disparity map between binocular images. In addition to bidirectional matching (BM), semi-global matching (SGM) [4] and graph cuts (GCs) [5], three common traditional algorithms, there are unsupervised depth prediction algorithms based on deep learning. The idea is to use a CNN based on an end-to-end framework to predict the disparity map corresponding to the image. Garg et al. [6] used the above idea to predict the depth map corresponding to the output image by training a CNN model based on the end-to-end framework.

Although the unsupervised learning method significantly im- proves predictive accuracy when compared to supervised learn- ing, there is still a disadvantage: that the unsupervised learning model convergence is based on the loss of image similarity, and the computation of image similarity is weak in the image gradient. The unsupervised model is not accurate enough for subtle depth differences in the prediction results. In contrast, the supervised learning model is based on the calculation of the loss of a single pixel, so it also has a better response to the difference in pixel points in the flat region.

Inspired by the research work in [6,7], which leverages the advantages and disadvantages of supervised learning and unsu- pervised learning, we proposes a semi-supervised depth predic- tion with binocular information and sparse laser data. This model inputs with binocular RGB images and Velodyne laser data from the KITTI data set, where binocular RGB images are first used for unsupervised learning to obtain disparity maps. It is then fused with sparse laser depth data as supervised learning for the final depth prediction. In the learning stage, we also introduce a small amount of sparse supervision information to limit the solution space, and the depth prediction accuracy of the final model is improved with real-time performance.

In addition, this paper improves the effectiveness of unsupervised depth with sparse depth information. In the experiment involving the accuracy prediction model on the KITTI data set, we obtain a linear mean square error of 4.211 and an accuracy ratio for the depth pixel of 86.2%, which is better than previous algorithms.

In the following paragraphs, Section 2 presents background and related work, Section 3 shows our architecture and algorithm, Section 4 explains the experiments and the final section offers the conclusion to this research.

2. Background and related work

2.1. Introduction to convolutional neural networks

In 1980, Kunihiko [8] first proposed the concept of the CNN, which is based on local connections between neurons and a hierarchical tissue image. CNN is mainly composed of a convolutional layer, a pooling layer, an activation function, a full connection layer, a softmax regression layer and other modules. The ordered connections of these modules constitute an end-to-end deep learning framework. In 1988, LeCun [9] proposed the LeNet-5 network structure, which used the gradient descent algorithm to train the CNN in order to classify handwritten numbers and obtained better experimental results than the conventional machine learning algorithm.

Since 2012, CNNs have developed rapidly, with AlexNet [10], ZFnet [11], VGNet [12] GoogleNet [13] and ResNet [14] models having been proposed in that order. Some applications, especially ImageNet [15] in 2012, and DeepFace [16] and DeepID [17,18] in 2013, can handle image classification and identification with large-scale database, which is not possible for conventional algorithms. Meanwhile, AlphaGo [19], based on a search tree and deep neural network, was developed by Google in 2016 and beat the human champion in the Go competition. These major events further established the important role of CNNs.

At present, CNNs are among the research hotspots in many scientific fields, especially in image recognition, image segmentation and image classification. Furthermore, they can also directly acquire the features used to predict the visual depth from the original image, while avoiding the complicated preprocessing of large-scale image sets, meaning that they will be enjoy widespread application in the near future.

2.2. Supervised depth prediction algorithm

One of the most classical models for depth prediction was proposed by the Eigen Group of New York University in 2014. This model divides the depth prediction into two steps. The first step is to predict the overall coarse-grained depth map, while the second step is to perform the fine-tuning depth estimation from previous coarse-grained information about the depth map results. At CVPR 2015, Laina et al. [2] proposed a depth prediction model based on the FCN [3] structure. As the fully connected layer is removed, the number of network parameters is reduced to accelerate convergence and, since the depth image is no longer limited to the fully connected layer, any size of the output map can be obtained in this framework. Chen et al. [20] changed the task from predicting fixed depth to relative depth with the relative relationship between predicted near and far pixels, greatly reducing the difficulty of the coefficient problem, making is more similar to visual perception in the human system.

2.3. Conventional unsupervised depth prediction algorithm

In binocular stereo vision, the restoration of three-dimensional (3D) information on the scene usually requires corresponding parallax information, and the depth information of each pixel point can be obtained by using a mathematical formula.

$$depth\,(I) = \frac{f * B}{dis\,(I)} \tag{1}$$

$$|d_{LR}(x, y) - d_{RL}(x + d_{LR}(x, y), y)| < th \tag{2}$$

In Eq. (1), B is the camera focal length variable, f is the binocular baseline length variable, $dis\,(I)$ is the parallax corresponding to the pixel in image I, and $depth(I)$ is the obtained image's 3D depth profile. Parallax calculation methods can be classified into two categories: conventional algorithms and those based on matching strategies such as BM, SGM and GCs.

BM takes the right image as a reference. First, the matching point in the reference picture, according to the matching cost, is found, then the matching point in the left image is found in the same way. If the matching point is found within the specified range from the target point, the match is successful.

SGM [4] is an improved version of the BM algorithm and constructs a disparity map by selecting the disparity value of each pixel point, then optimizes the disparity value of each pixel point by minimizing the global energy function of the disparity map. SGM optimizes the mismatch problem caused by occlusion in the BM algorithm by matching uniqueness detection, as well as removes residual noise after LR and unique detection via connected region detection. The optimization effect is shown in Fig. 1b-c. In the case of GCs [5], an optimization algorithm is applied to solves the problem of energy minimization in the image field. It can be used for binocular disparity estimation. The effect is shown in Fig. 1d.

Comparing the results of the three conventional methods, it can be found that: the edge distortion of the object in the BM algorithm is severe and has more noise points; the SGM algorithm optimizes the shortcomings of the BM, such that the edge information in the result image can be found to be optimized; and the GC algorithm results are the best of the three, with clear edge information and a few mismatching points, although it is also the most time-consuming algorithm of the three.

2.4. Unsupervised depth prediction with a deep learning algorithm

In addition to the above three common conventional algorithms, there are also unsupervised depth prediction approaches based on deep learning algorithms. The idea is to use the CNN based on an end-to-end framework to predict the disparity map of the image. The core of the unsupervised algorithm is the disparity map between binocular images. If the binocular disparity map $dis\,(I)$ is known, then the image I_r corresponding to the right eye can be generated by the left eye image I_l and the disparity map, which can be calculated by Eq. (3). In the same way, the left eye image can also be generated by the right eye image. The unsupervised depth prediction model utilizes this binocular image constraint. In the network model training process, the left eye image I_l predicts the disparity map d_{pred}via the network model $F\,(I)$, then maps the right eye image to the left eye's estimated image $I_l^{\wedge}$ according to the disparity map. If the predicted disparity map is highly accurate, then $I_l^{\wedge}$ and I_l will be very similar. In the absence of depth information supervision, the model has to learn how to minimize the difference between the estimated leftimage and the original left image.

$$I_r = I_l\,(dis\,(x_l) + x_l) \tag{3}$$

Garg et al. [6] presented a typical depth prediction model, which uses the above idea to predict the depth map corresponding to the output image by training a CNN model based on the end-to-end framework. The network model is shown in Fig. 2. The forward process of the model is to generate a warp image by fusing the right image and to predict the inverse depth from the CNN processing of the left image. This reverse process is to minimize the gap between the reconstructed image and the original left image as a reconstruction error of the model.

After the unsupervised prediction algorithm proposed by Garg et al. an improved version of the model is proposed in [21]. Since the relationship between the binocular images is mutual, the right image can be reconstructed according to the drawing as well as reconstructed. Further, the practice of Garg et al. only utilizes the constraint between the original left image and the reconstructed left image. Godard [7] creatively uses a network model to predict the disparity maps for the left and right images and then reconstruct the right and left images. Not only that, because the predicted left and right disparity maps also have a binocular geometric parallax relationship, reconstructed left and right disparity maps can also be obtained.

This transforms the single left image constraint in the unsupervised learning process into a left image and a right image, along with their disparity maps. In addition to adding constraints to the unsupervised reverse process, Godard [7] also adjusted the network architecture. In the upsampling process, the disparity estimation map corresponding to the current feature is increased by interpolation and passed to the next upsampling module. This is similar to coarse-grained and fine-grained fine-tuning in each upsampling, which makes the predictions more accurate.

3. Architecture and algorithms

At present, the research results based on deep learning can be divided into two categories according to different learning strategies: (1) supervised learning, by training the network model in the RGBD image data set, obtains a CNN model for predicting the depth of each pixel from a single image; (2) unsupervised learning, without real depth maps, automatically learns a model capable of predicting disparity maps using the intrinsic constraints between binocular images.

The most effective method for supervised learning in KITTI is the method in [1], with a prediction result of $\delta_1 = 69.2\%$. The optimal method for unsupervised learning is the method in [21], and the prediction result is $\delta_1 = 87.3\%$. In our analysis, we found the issues of previous works to be:

(1) The KITTI data set is derived from an unmanned mission, the image acquisition scene is mainly outdoors, and the depth information is acquired using 64-line laser radar; thus, the sparse real depth map cannot generate a suitable model for convergence in order to obtain a better local minimum.

(2) Unsupervised learning uses binocular image information from the data set. Since there is a large number of pixel pairs between binocular images that comply with parallax geometry from the stereo vision, convergence is not performed well enough to obtain a better local minimum from the binocular image.

In summary, although the unsupervised learning method represents a significant improvement in predictive accuracy, compared to supervised learning, it can be seen from the above description that the unsupervised model involves low resolution for pixels with subtle depth differences in the prediction results. In contrast, the supervised learning model is based on the loss computation function from pixels of a single image, and offers better response to the difference in pixel points in a flat region.

3.1. Proposed architecture

Inspired by the research work in [6,7], we leverage the advantages of supervised learning and unsupervised learning strategies, then propose a semi-supervised depth prediction model using binocular information and sparse laser data. The overall architecture is shown as Fig. 3.

The model is input with binocular RGB images and Velodyne laser data in the KITTI data set, where binocular RGB images are used for unsupervised learning to obtain disparity maps, while sparse laser depth data are also used to supervise the depth prediction results.

The end-to-end CNN model in this paper is an improved encoder–decoder framework, in which the new idea of multi-loss and sub-pixel rearrangement is applied to the upsampling module in this architecture. The detailed frame structure is shown in Fig. 4.

For the parameter setting of the CNN model, the feature ex- traction of the algorithm is based on the encoder part $f\ (x)$, using the ImageNet-based pretraining model, ResNet-50. Based on the network model, the fully connected layer is removed and the final output of the network is characterized as an $H \times W \times 2048$ feature matrix. The input to the decoder part $g\ (x)$ is parsed as a feature. The decoder part $g\ (x)$ of the end-to-end framework is responsible for mapping the obtained feature map from the feature space onto the depth space, such that a dense depth map is finally obtained.

3.2. Loss function

A new loss of the model consists of unsupervised and supervised losses is proposed

The loss of the model is the difference between the reconstructed image and the original image. The more accurate the parallax result predicted by the model, the smaller the difference between the reconstructed image and the original image. The learning process of the model is the process of continuously reducing the difference between the reconstructed image and the original image. According to the literature published by Nvidia and the MIT [21], in the field of image processing based on deep learning, commonly used indicators to measure the difference between images are as follows: L1, L2, PSNR and SSIM [22].

The PSNR (peak signal-to-noise ratio) is the most widely used image quality evaluation index; but, as it is based on the error of the pixel in the corresponding position, this leads to an evaluation result which is often inconsistent with the subjective feeling of the person. At the same time, due to the point-to-point error calculation method, the PSNR does not take into account the correlation between the pixel points in the image and other pixels in the field. The above characteristics mean that the PSNR is commonly used for quality evaluation before and after image compression.

Compared with the PSNR, the SSIM (structural similarity index) is a quality evaluation index that calculates the pixel correlation in the image domain, which is more suitable for scenes with distortion. As the SSIM features highly in images from nature images, it can measure the quality of the image in terms of the brightness, contrast and structural information in the image, as shown in Fig. 5.

The computation of the SSIM can be divided into two stages: (1) computation of the luminance evaluation index $l\,(x,\ y)$ based on the image mean values μ_x and μ_y, the contrast evaluation index $c\,(x,\ y)$ based on image variances σ_x and σ_y, and the structural correlation index $s\,(x,\ y)$ based on image variances σ_x and σ_y, according to Eqs. (4)–(6); (2) the final quality evaluation index is obtained from Eq. (7), in which $\alpha = \beta = \gamma = 1$, $k_1 = 0.01$, $k_2 = 0.01$, $L = 255$ are often used. The range of the SSIM is 0–1, and the distortion degree is inversely proportional to the value.

$$l\,(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{4}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{5}$$

$$s\,(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{6}$$

$$SSIM(x, y) = [l(x, y)]^{\alpha}[c(x, y)]^{\beta}[s(x, y)]^{\gamma} \tag{7}$$

where C_1, C_2, C_3 are determined as follows:

$$C_1 = (k_1L)^2\,,\ C_2 = (k_1L)^2\,,\ C_3 = C_2/2 \tag{8}$$

The loss function of the model consists of two parts: an unsupervised loss from binocular information and a supervised loss from laser sparse depth. Deriving the decoder from the Godard [7] model, our improved model first predicts the binocular disparity maps $disp_l^{\wedge}$ and $disp_r^{\wedge}$ from the left image, then obtains the reconstructed left image I_l and right image I_r with the disparity estimation map. This is similar to images I_l and I_r in Eq. (3). Next, the reconstructed maps $disp'_l$ and $disp'_r$ are generated according to the left and right parallax maps. Finally, the depth prediction map $disp_l^{\wedge}$ is obtained according to the left disparity estimation map $depth_l^{\wedge}$. Based on the above intermediate results, the final loss of the model is obtained according to the semi-supervised model with loss function architecture, as shown in Fig. 6. The loss before and after image reconstruction is a combination of L1 and SSIM loss function.

The network model in this paper uses an end-to-end framework, where the encoder uses ResNet50, which removes the fully connected layer, and the decoder consists of a series of upsampling modules. Between the output disparity map and the depth, a convolution layer is added in order to learn the mapping relationship between parallax and real depth. The loss of the model consists of unsupervised reconstruction losses and supervisory losses, where the reconstruction losses are obtained from Eq. (9) and the supervisory losses are obtained from Eq. (10). Before model training, this paper uses the trained model parameters based on ImageNet in order to initialize the parameters in the model. For the parameters of the decoder module, initialization is made by referring to the MSRA method described in [23]. In the model training process, the learning rate for the feature extraction parameters is reduced to one tenth of the global learning rate, while the upper sampling module learning rate uses the global learning rate. In addition, according to different learning stages, the global learning rate is correspondingly attenuated. When the model loss is small, the global learning rate is attenuated to one tenth.

$$C_{rc}^{U} = \frac{1}{n}\sum_{i,j}\alpha\,\frac{1 - SSIM\left(I_{ij}, I_{ij}^{\wedge}\right)}{2} + (1-\alpha)\left\|I_{ij}, I_{ij}^{\wedge}\right\|_1 \tag{9}$$

$$C_{l}^{S} = \frac{1}{n}\sum_{i,j}\left|\left|dp_{ij} - dp_{ij}^{\wedge}\right|\right| \tag{10}$$

4. Experiments

4.1. Experimental data set

In order to verify the effectiveness of the feature parsing module proposed in this paper, we use the KITTI data set, which was jointly established by the Karlsruhe Institute of Technology and the Toyota American Technology Research Institute. It is the largest data set in the field of computer vision for autopilot applications. The data set can be used to evaluate the performance of stereo vision, visual tracking, visual odometry, 3D object detection and other applications in the vehicle environment. It also contains real-world scenes for specific driving scenarios, such as urban, rural, campus and highway scenes.

The KITTI data set (url: http://www.cvlibs.net/datasets/kitti/) contains, for example, image data, laser data and GPS data. The data collection platform is shown in Fig. 7a. The platform is equipped with two grayscale cameras, two color cameras, one 64-line 3D laser radar and a GPS navigation system. In this paper, binocular RGB image data and laser radar data in the KITTI data set are used for the experiment. The RGB image and the laser sparse depth information are shown in Fig. 7b.

4.2. Experimental setup and experimental environment

In order to verify the effectiveness of the fusion model with the accuracy of depth prediction, we set up experiments based on the following two comparative groups: (1) using binocular data and 64-line laser depth data to train the network model, then verifying the effectiveness by adding sparse information as a comparison with existing unsupervised algorithms; (2) extracting sparse depth data from the corresponding depth information as four-, eight-, 16-, 32- and 64-line laser scanning. The improved model was trained and evaluated, and the influence of different laser line numbers on depth prediction was also evaluated.

This paper uses the KITTI image data set as experimental data. The data set is divided according to the partitioning method for depth prediction sub-tasks. It contains 86,000 frames of image data: 43,000 image frames comprise the training set, while the test set has 34,000 image frames with corresponding laser scanning data. In the model for the training process, we employ data augmentation to increase the data set which uses the following methods: horizontal flip image, gamma adjustment, random adjustment of color channels, and contrast- and brightness-adjusted methods.

The laser sampling method is different from the random sampling method with the KITTI experimental data set as in [24]. Our method extracts the laser scanning of plane data as four, eight, 16 and 32 lines according to the structure information from 64-line laser radar.

According to the Velodyne HDL-64e user manual, the longitudinal scanning angle of the lidar is $-24.8^{o}\sim +2.0^{o}$, as shown in Fig. 8 (left). The spatial laser scan involves a space rotation around 360^{o}, and the laser point set in the experimental data set is distributed in this lidar space. Since the longitudinal angle of each line is the same, the lidar space is divided into 64 sub-spaces with an angle of 0.42^{o}, and the real depth information required for the experiment is extracted from 64 sub-spaces with a fixed step size. The laser sampling result is shown in Fig. 8 (right).

In the experiment, we conduct a comparison of the depth pixel ratio with different laser sampling lines. The results are shown in Table 1, which shows that the true depth information of the sampled KITTI data set is very sparse.

Table 1
Pixel ratios with the known depth of the sample results.

Number of laser lines	4	8	16	32	64
Depth pixel ratio	1.40	2.70	4.50	9.70	19.60

Table 2
Comparison of improved unsupervised algorithms and mainstream algorithms.

	RMSE	Rel	δ_1	δ_2	δ_3
Eigen et al. [1]	7.216	0.228	67.9	89.7	96.7
Garg et al. [6]	5.104	0.169	74.0	90.4	96.2
Godard et al. [7]	5.849	0.141	81.8	92.9	96.6
Godard et al. [7] + CS	5.763	0.136	83.6	93.5	96.8
Liu et al. [25]	6.523	0.275	67.8	89.5	96.5
Zhou et al. [26]	6.565	0.275	71.8	90.1	96.0
Mahjourian et al. [26]	6.220	0.250	76.2	91.6	96.8
Yin et al. [27]	5.737	0.232	84.6	93.4	97.2
Ours (with 64 lasers)	4.211	0.124	86.2	94.3	97.4

4.3. Comparison

In order to confirm the addition of sparse laser information, which can effectively improve the accuracy of depth prediction, we perform a comparative experiment with other mainstream algorithms. In this experiment, a binocular image and laser depth information are used from the KITTI data set. In this comparison, each algorithm is performed 20 times on the training data set, with a batch size of 20. The final evaluation results are shown in Table 2.

It can be seen from the experimental results that the addition of sparse real laser depth information makes a significant improvement to prediction accuracy. Compared with the existing supervised depth prediction algorithm [1], the improved unsupervised depth prediction algorithm reduces the root mean square error (RMSE) by 41.6%, and the depth prediction accuracy from 67.9% to 86.2%. Compared with the existing unsupervised prediction algorithm [7] (without the Cityscapes data set), the RMSE is reduced by 1.638, and the depth prediction accuracy δ_1 is increased by 5.6%.

From the experimental comparison, we can see that the unsupervised algorithm significantly improves prediction accuracy compared to the current mainstream supervised algorithm, which is due to the addition of the intrinsic constraint between binocular images. At the same time, compared with the supervised algorithm, the sparse laser depth information makes the model prediction closer to the real distribution. This validates the effectiveness of the added sparse laser depth information to improve the accuracy of final depth prediction. The depth prediction examples of our improved model are shown in Fig. 9.

It can be seen from the improved unsupervised depth prediction algorithm renderings that the depth prediction results already contain obvious object contour information, such as vehicles, pedestrians, plants on both sides of the road, and road signs.

For the processing speed of this algorithm, the module with the largest amount of computation is the depth prediction of CNN, and this module has only the operation of convolution, and no function of neural network. Therefore, our algorithm will be faster than the normal CNN algorithm for image classification and object detection. According to our experiments, the inference time of our algorithm is less than 0.05 s per frame, and it is more suitable than real-time processing.

To further quantify the impact of sparse real depth information on unsupervised predictive models, we designed a second set of comparative experiments. We extracted the sparse depth data corresponding to the four-, eight-, 16-, 32- and 64-line lasers as supervised information and added them to the model for model training and evaluation. The final prediction results are shown in Table 3.

It can be seen from the experimental results of lasers with different line numbers that, as the linearity of the number of laser lines increases, the accuracy of depth prediction tends to increase as well. This indicates that the sparse supervision information is effective for improving the prediction accuracy of the model.

Table 3
The effect of different laser line numbers on the prediction results.

	RMSE	Rel	δ_1	δ_2	δ_3
Exp. 1 (4 lines)	5.104	0.159	84.1	93.4	97.2
Exp. 2 (8 lines)	4.976	0.148	84.7	93.8	97.2
Exp. 3 (16 lines)	4.685	0.134	85.4	93.8	97.3
Exp. 4 (32 lines)	4.519	0.126	85.9	94.1	97.4
Exp. 5 (64 lines)	4.211	0.124	86.2	94.3	97.4

4.4. Applied experiment

This research work is suitable for supporting the work of the unmanned sweeping vehicle project. The unmanned sweeping vehicle platform is shown in Fig. 10 (a), which is equipped with a binocular camera, 16-line laser radar and GPS equipment. The depth prediction algorithm, based on our improved method, plays an important role in obtaining 3D structural information on the environment and enabling high-level semantic segmentation for path planning.

In this application, the trained model with the KITTI data set can be applied to our university scene as well. The depth prediction results are shown in Fig. 10b. From the prediction results, clearer contour information of the object can be obtained and the predicted depth image can reflect the true 3D structure as well.

5. Conclusion

This paper first introduces the unsupervised depth prediction algorithm with the binocular parallax images. Further, we analyze the advantages and disadvantages of supervised learning and unsupervised learning. Then, an improved semi-supervised depth prediction model combining these two learning strategies is proposed. The improved model utilizes the unsupervised prediction of binocular images by introducing sparse laser depth information. Finally, based on the KITTI data set, two sets of comparative experiments are carried out to confirm the effectiveness of other mainstream models. In addition, the depth prediction experiment also considers the influence of different numbers of laser lines. Finally, the effect of the improved algorithm is demonstrated on our unmanned sweeping platform, with depth rendering images from our university scene.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like acknowledge the Shenzhen Government for supporting Project Numbers (KQJSCX20170726104033357, JCYJ20150513151706567 and JCYJ20160531191837793). Furthermore, this work was partially supported by a grant from the Department of Industrial and Systems Engineering at the Hong Kong Polytechnic University (H-ZG3K).

References

[1] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, in: Neural Information Processing Systems, Montreal, 2014, pp. 2366–2374.

[2] I. Laina, C. Rupprecht, V. Belagiannis, et al., Deeper depth prediction with fully convolutional residual networks, in: International Conference on 3D Vision, 2016, pp. 239–248.

[3] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Computer Vision and Pattern Recognition, Boston, 2015, p. p 3431–3440.

[4] H. Hirschmuller, Accurate and efficient stereo processing by semi-global matching and mutual information, in: IEEE Computer Vision and Pattern Recognition, San Diego, 2005, pp. 807–814.

[5] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans. Pattern Anal. Mach. Intell. 23 (11) (2001) 1222–1239.

[6] R. Garg, V.K. BG, G. Carneiro, et al., Unsupervised CNN for single view depth estimation: geometry to the rescue, in: European Conference on Computer Vision, Amsterdam, 2016, pp. 740–756.

[7] C. Godard, O.M. Aodha, G.J. Brostow, et al., Unsupervised monocular depth estimation with left-right consistency, in: IEEE Computer Vision and Pattern Recognition, Honolulu, 2017, pp. 6602–6611.

[8] K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybernet. 36 (4) (1980) 193–202.

[9] Y. Lecun, L. Bottou, Y. Bengio, et al., Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

[10] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: International Conference on Neural Information Processing Systems, Curran Associates Inc., 2012, pp. 1097–1105.

[11] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, 8689 (2013) 818–833.

[12] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Sci. (2014).

[13] C. Szegedy, W. Liu, Y. Jia, et al., Going deeper with convolutions, 2014, pp. 1–9.

[14] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, in: Computer Vision and Pattern Recognition, IEEE, 2016, pp. 770–778.

[15] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: International Conference on Neural Information Processing Systems, Curran Associates Inc., 2012, pp. 1097–1105.

[16] Y. Taigman, M. Yang, M. Ranzato, et al., Deepface: Closing the gap to human-level performance in face verification, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2014, pp. 1701–1708.

[17] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting10, 000 classes, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2014, pp. 1891–1898.

[18] Y. Sun, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, 27, 2014, pp. 1988–1996.

[19] D. Silver, A. Huang, C.J. Maddison, et al., Mastering the game of go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489.

[20] W. Chen, Z. Fu, D. Yang, et al., Single-image depth perception in the wild, in: Neural Information Processing Systems, Barcelona, 2016, pp. 730–738.

[21] F. Tschopp, J.N.P. Martel, S.C. Turaga, et al., Efficient convolutional neural networks for pixelwise classification on heterogeneous hardware systems, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI 2016), IEEE, 2016.

[22] Z. Wang, A.C. Bovik, H.R. Sheikh, et al., Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[23] K. He, X. Zhang, S. Ren, et al., Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: IEEE International Conference on Computer Vision, Boston, 2015, pp. 1026–1034.

[24] F. Ma, S. Karaman, Sparse-to-dense: Depth prediction from sparse depth samples and a single image, 2017, arXiv preprint: arXiv:1709.07492.

[25] T. Zhou, M. Brown, N. Snavely, et al., Unsupervised learning of depth and ego-motion from video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1851–1858.

[26] R. Mahjourian, M. Wicke, A. Angelova, Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5667–5675.

[27] Z. Yin, J. Shi, Geonet: Unsupervised learning of dense depth, optical flow and camera pose, in: Proceedings of the IEEE Conference on Computer

[28] Vision and Pattern Recognition, 2018, pp. 198

[29] 3–1992.

Fig. 1. The effect comparison of conventional algorithms: (a) the original image, (b) the results of the BM algorithm, (c) the results of the SGM algorithm, (d) the results of the GC algorithm.

Fig. 2. Data flow of the conventional depth prediction model.

Fig. 3. Algorithmic framework for semi-supervised image depth prediction.

Fig. 4. End-to-end CNN model.

Fig. 5. SSIM measurement for loss function.

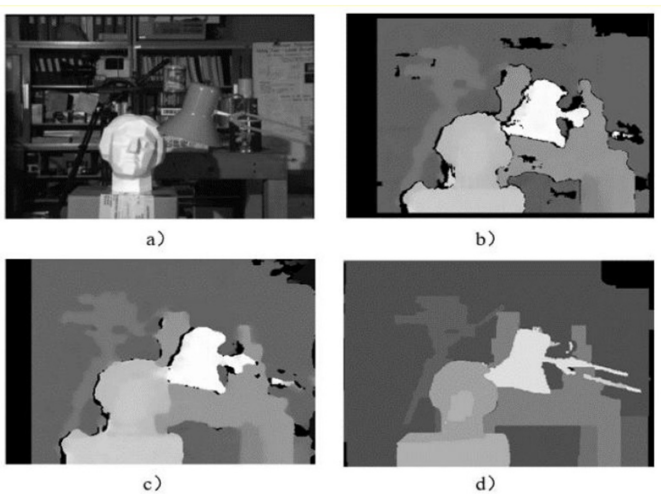
Fig. 6. Composition of semi-supervised model loss.

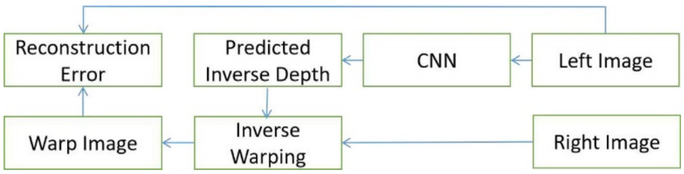
Fig. 7. (a) KITTI data acquisition platform, (b) KITTI RGB image (top) and laser sparse depth information (bottom) (from http://www.cvlibs.net/datasets/kitti/).

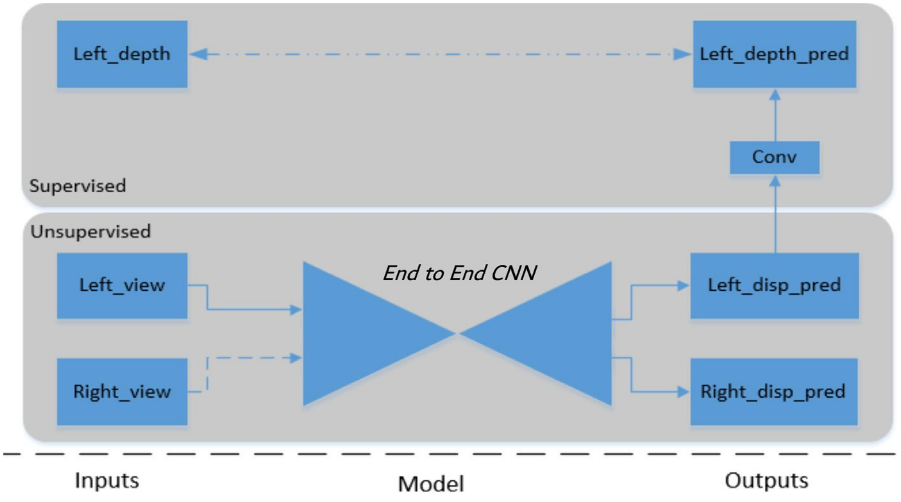
Fig. 8. Laser sampling method and effect.

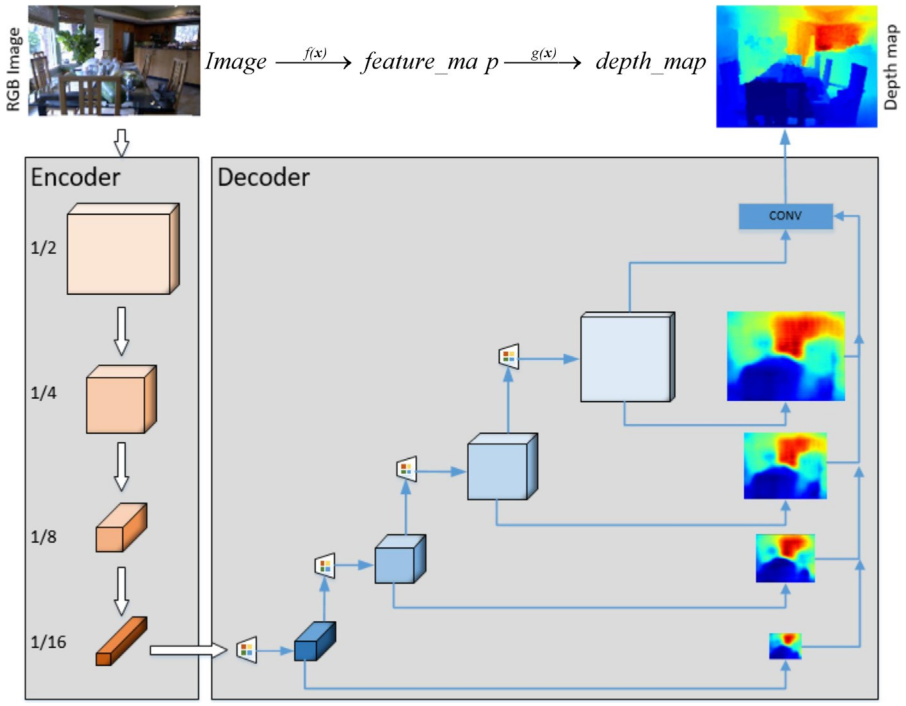
Fig. 9. Improved model for depth rendering image prediction.

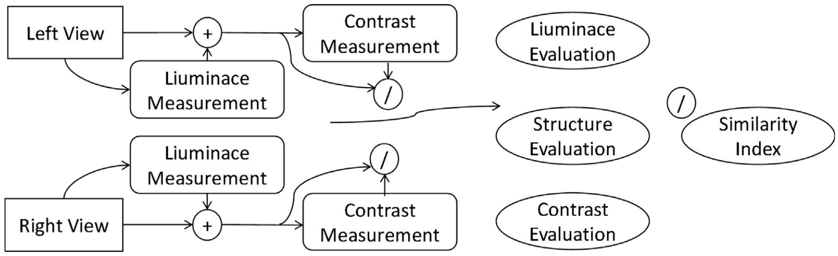
Fig. 10. (a) Unmanned sweeping vehicle platform, (b) depth rendering image of our university scene.

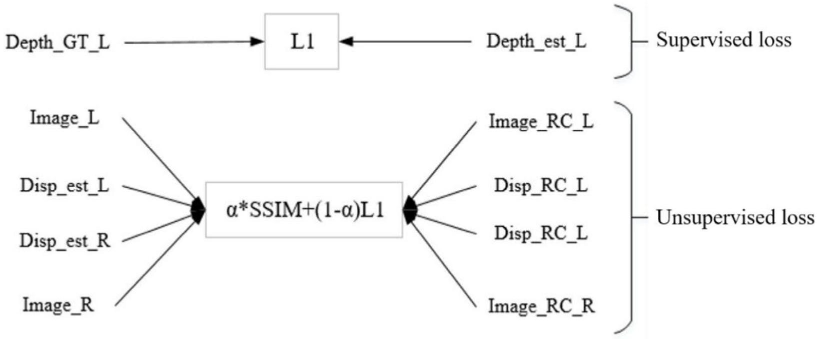

a)
b)
c)
d)


Reconstruction Error
Predicted Inverse Depth
CNN
Left Image
Warp Image
Inverse Warping
Right Image

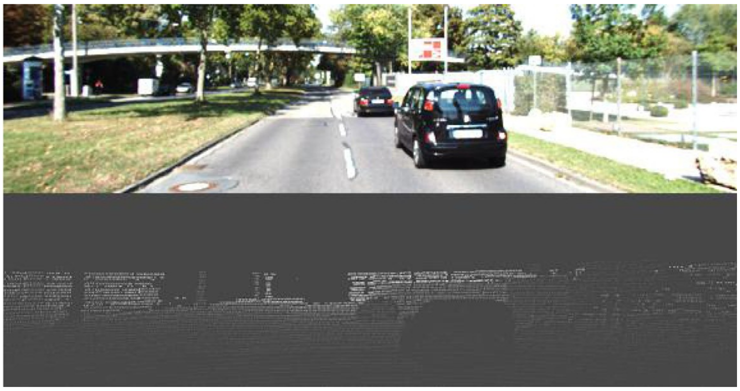

Left_depth
Left_depth_pred
Conv
Supervised
Unsupervised
End to End CNN
Left_view
Right_view
Left_disp_pred
Right_disp_pred
Inputs
Model
Outputs

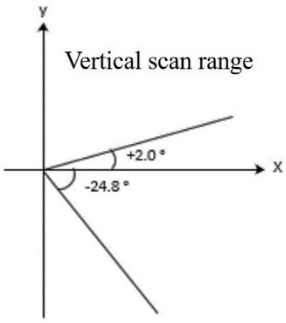

RGB Image
Image $\xrightarrow{f(x)}$ feature_ma p $\xrightarrow{g(x)}$ depth_map
Depth map
Encoder
Decoder
1/2
1/4
1/8
1/16
CONV


Left View
+
Contrast Measurement
Liuminace Measurement
/
Liuminace Evaluation
Structure Evaluation
/
Similarity Index
Contrast Evaluation
Liuminace Measurement
/
Right View
+
Contrast Measurement


Depth_GT_L
L1
Depth_est_L
Supervised loss
Image_L
Disp_est_L
Disp_est_R
Image_R
α*SSIM+(1-α)L1
Image_RC_L
Disp_RC_L
Disp_RC_L
Image_RC_R
Unsupervised loss

Velodyne HDL-64E Laserscanner
Point Gray Flea 2 Video Cameras
OXTS RT 3003 GPS / IMU





a) Left eye image b) Depth prediction result