Unsupervised Monocular Depth Estimation with Scale Unification

Xinyan Jiang[1], Meng Ding[2]
College of Civil Aviation
Nanjing University of Aeronautics and Astronautics
Nanjing, China
jiangxy@nuaa.edu.cn, nuaa_dm@nuaa.edu.cn

Abstract—**Due to the importance of high precision depth data for driving assistance system, related methods to predict scene depth without ground truth have attracted more attention than ever. Therefore, this paper comes up with a novel unsupervised monocular depth estimation model. Based on the structure of Deep3D, the proposed model reduces the influence of the diverse scale of targets by processing input images as pyramid structure. Furthermore, the proposed model improves the 'hole' phenomenon in depth maps and solves it by up-sampling the intermediate disparity maps to the initial resolution. The experimental results show that this model achieves convincingly accuracy on classical KITTI dataset and excellent generalization capability on images randomly captured in real driving scene.**

Keywords-monocular depth estimation; multi-scale; driving assistance system; unsupervised learning

I. Introduction

After decades of research, vision-based driving assistant system [1] has popularized in the field of passenger cars and trucks. Among all terms of visual information, depth information plays a quite important role in driving assistant system. For example, the collision avoidance system gives out collision warnings by calculating the depth information between obstacles and the vehicle. When the distance between pedestrians and the vehicle is too small, the pedestrian protection system will automatically take measures to decelerate the vehicle. In such case, only when the depth information is obtained, can the driving assistant system get the connection with external environment accurately, so that the warning subsystems can function normally.

Currently, diverse sensors can be used to obtain the depth information, such as 2-D and 3-D LiDAR. However, soaring cost and limited installation condition greatly restrict their application. Therefore, researchers have to explore approaches to restore 3-D structural information of scenes from mere images.

Traditional visual methods include stereo matching [2], [3], which uses two cameras to observe the same scene so that we obtain two calibrated images. After that, depth information would be calculated from disparity between two images by triangulation. However, this method requires at least two cameras and their relative positions must remain fixed, which limits its usage. On the contrary, the depth estimation method based on monocular vision has lower requirements on the camera and thus has a broader application prospect.

Most previous monocular depth estimation methods, for example, structure from motion [4], defocus cues [5],[13] and illumination variation [6], are based on classic geometric constraints. Recently, with the success of convolutional neural network on other visual tasks, many researchers begin to explore the application of deep learning to solve the challenge of image depth estimation [7]. By making use of the powerful learning ability of neural networks, people can design various models to dig the relationship between original images and depth maps [8]. There are mainly two types of methods based on deep learning, namely the supervised learning methods and the unsupervised ones. The supervised methods rely on the ground truth of image depth, which is not always available. For this reason, more and more attention has been attracted by the unsupervised methods. However, existing works using unsupervised learning cannot solve the problem of detail blurring in low-scale feature maps. To deal with that, we propose a novel method which samples all the depth maps to the size of input so as to avoid voids in the depth maps.

II. Related Work

Unsupervised depth estimation refers to the problem of estimating depth in absence of ground truth information. In this case, we usually adopt other supervisory to complete the training process. The most common one is the constraints during image reconstruction.

Some researches use monocular videos to reconstruct images. In 2017, Zhou et al. [10] used geometric information from different perspectives of continuous frames as supervision to estimate both depth and pose. Influenced by direct visual odometry in 2018, Wang C presented a direct method to obtain unsupervised monocular depth estimation. The difficulty of this self-supervised method is that not only the scene depth but also the camera pose between frames should be estimated.

Another representative supervisory signal comes from the spatial constraint of stereo image pairs. Xie et al. [11] proposed Deep3D network in 2016. This network solved the problem of synthesizing stereo image pair from a single image and provided a basis for such kind of method. Garg et al. [12] generated the depth map through the full convolutional neural network in the encoding stage. Then they reconstructed the input image using the traditional

978-1-7281-4653-9/19/$31.00 ©2019 IEEE
DOI 10.1109/ISCID.2019.00072

binocular camera measuring principle in the decoding stage. The loss function works using the difference between the reconstructed images and the original inputs. In 2017, Godard [13] et al. used binocular images for image reconstruction and generated disparities during training. Then they restored the depth according to the corresponding disparity map at the test time. But they didn't address the issue that depth maps are partially blurred. To solve this problem, we come up with a scale unified method.

III. Method

Without ground truth depth, our method takes depth estimation as a problem of visual synthesis. This is implemented by using intermediate disparities to constrain the network. Then the disparity maps are extracted from the model to restore depth. Finally, we sample all intermediate depth maps to the same scale and improve the blur in the predicted depth maps.

A. Image Reconstruction

Assuming that a single image $\boldsymbol{I}$ is the input at the test period, we need to predict depth of the scene $\hat{\boldsymbol{d}}$. Inspired by the method of Godard, we transform this problem into image reconstruction, and generate disparities from each perspective simultaneously during training, to improve the reconstruction accuracy.

During training, $\boldsymbol{I}^l$ and $\boldsymbol{I}^r$ pictured from the same scene can be obtained from the left and right cameras. When using $\boldsymbol{I}^l$ as input, the network predicts $\boldsymbol{d}^r$ pixel by pixel which is then applied to t $\boldsymbol{I}^l$ to reconstruct $\hat{\boldsymbol{I}}^r$.

When $\boldsymbol{I}^l$ being the only input image, we can generate two disparity maps, $\boldsymbol{d}^l$ and $\boldsymbol{d}^r$, at the same time (Fig. 1). After that, we make use of $\boldsymbol{I}^l$ and $\boldsymbol{d}^r$ to reconstruct $\hat{\boldsymbol{I}}^r$, the same situation happens among $\boldsymbol{I}^r$, $\boldsymbol{d}^l$ and $\hat{\boldsymbol{I}}^l$. Then, the reconstructed images of two different perspectives are compared with the input left images and right ones respectively, where the appearance loss is used to constrain the error to improve the reconstruction accuracy.

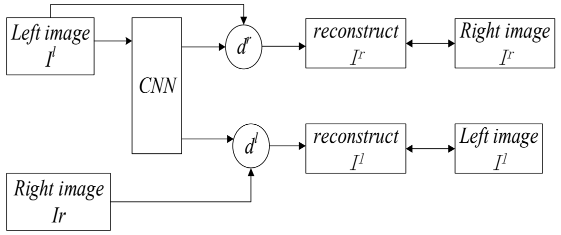

Figure 1. Our approach to produce disparities for both images

B. Loss Function

As is mentioned above, appearance loss C_a is required during reconstruction period. Taken $\boldsymbol{I}^l$ and the reconstructed $\hat{\boldsymbol{I}}^l$ as an example, C_a is composed of structural similarity index (S) and L1 loss function. S is an index to measures structural similarity between two images. And L1 loss is less sensitive to outliers.

$$C_a = \frac{1}{K}\sum_{i,j}\alpha\frac{1-S(\boldsymbol{I}_{ij}^l - \hat{\boldsymbol{I}}_{ij}^l)}{2} + (1-\alpha)\left\|\boldsymbol{I}_{ij}^l - \hat{\boldsymbol{I}}_{ij}^l\right\|. (1)$$

With left-right consistency loss, C_l, disparity map of each view can be calculated through corresponding disparities and disparity map.

$$C_l = \frac{1}{N}\sum_{i,j}\left|\boldsymbol{d}_{ij}^l - \boldsymbol{d}_{ij+d_{ij}^l}^r\right|. \quad (2)$$

What's more, there is a depth smoothing constraint term C_d in the loss function.

$$C_d = \frac{1}{N}\sum_{i,j}\left|\partial_x \boldsymbol{d}_{ij}^l\right| e^{-\left\|\partial_x \boldsymbol{I}_{ij}^l\right\|} + \left|\partial_y \boldsymbol{d}_{ij}^l\right| e^{-\left\|\partial_y \boldsymbol{I}_{ij}^l\right\|}. \quad (3)$$

Each loss can be calculated both in the left images and the right ones using the same formula. The final loss function is composed of three items:

$$C = C_a^l + C_a^r + C_d^l + C_d^r + C_l^l + C_l^r. \quad (4)$$

C. Multi-scale Unification

When we reconstruct images during training, we use bilinear sampler [16] to provide spatial invariance. However, bilinear sampler may cause the training target falling into local minimization. To prevent this situation, existing model resize the input image into four scales. The final loss is the sum of the loss at each scale. However, when dealing with lower-scale depth maps, there usually exists errors like artifacts and hole phenomenon in depth maps. In low resolution regions, the texture is often ambiguous, resulting in bigger appearance loss. In this case, holes may appear in the final depth map. For solving this difficulty, we put forward a method to separate the resolutions [18] of disparity maps from that of images (Fig. 2).

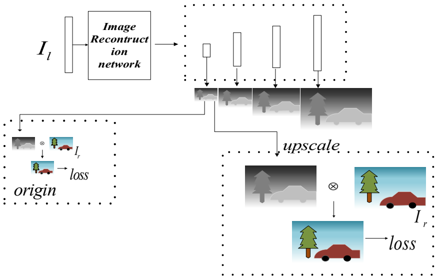

Figure 2. Steps for scale unification

TABLE I. Results on KITTI dataset

Methods	Supervised	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
		Lower is better				Higher is better		
Eigen et al.	Yes	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Zhou et al.	No	0.183	1.565	6.709	0.270	0.734	0.902	0.959
Godard et al.	No	0.128	1.183	**5.517**	**0.279**	0.815	0.922	0.968
Ours	No	**0.116**	**1.164**	5.562	0.284	**0.828**	**0.926**	**0.970**

During my experiment, we sample all disparity maps up to the resolution as same as input image and compute the appearance loss at this scale. After that, the image was reconstructed through the corresponding original image and corresponding disparity maps. By using this method, the disparity maps at each resolution were constrained to restore the image at the same scale of the input one.

D. *Network Architecture*

We adopt the encoder-decoder architecture similar to Deep3D network. During the encoding stage, a ResNet-50 network with three-layer identity residual blocks is deployed as the feature extraction model (Fig. 3).

Firstly, taking into account the objects of various sizes, input images are resized to different resolutions to form a pyramid structure [15]. Then these images are sent into the model for feature extraction. Secondly, the decoding network takes the extracted features as input and complete the reconstruction. During this process, we use skip connections to concatenate the feature map in the encoding process with the corresponding map in deconvolution layers. Thirdly, the disparity maps obtained in the process of decoding are up-sampled to the input scale.

IV. Experiments

Our experiments are completed at the desktop workstation with the graphics card of NVIDIA GeForce GTX 1080Ti and the training system of Ubuntu 14.04.

In this section, our model is made comparison with other methods of depth estimation, including supervised and unsupervised methods. We not only show the predicted depth maps, but also evaluate the results by calculating the metrics of Abs Rel et al., which both prove the superiority of our method.

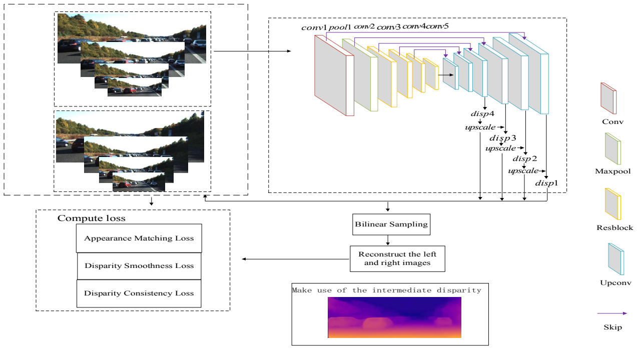
conv1 pool1 conv2 conv3 conv4conv5
disp4
upscale
disp3
upscale
disp2
upscale
disp1
Conv
Maxpool
Resblock
Upconv
Skip
Compute loss
Appearance Matching Loss
Disparity Smoothness Loss
Disparity Consistency Loss
Bilinear Sampling
Reconstruct the left and right images
Make use of the intermediate disparity

Figure 3. Our network architecture

A. *Details*

We implement the depth estimation model in TensorFlow, and use the KITTI dataset with stereo image pairs as the training set, where 29000 pairs of images are taken as training dataset and 2900 pairs as validation dataset. During training, we set the basic learning rate of 0.0001. The total training period lasts 70 training epochs. It costs us 34 hours to complete the entire training process.

B. *Analysis of Experiment Results*

We firstly take a comparative experiment on KITTI dataset. In order to show the results of depth estimation vividly, we present the predicted depth maps of different methods (Fig. 4). The first method (c) is supervised, which uses ground truth depth when training. The second one (d) is unsupervised depth estimation method using monocular video as input by Zhou et al. Other methods (e) and (f), including ours (f), use stereo pairs when training and only left image when testing.

Besides, we present the metrics results of these depth estimation algorithms mentioned above (Table I). By comparing the data in the first and fourth rows, we can find out that although our model does not use ground truth as supervision, it is still feasible and can achieve superior results. Through the comparison between the data in the second, third and fourth rows, we can see that our scale unification method can effectively ameliorate the results of image restoration and accuracy of depth estimation. Moreover, to show the generalization capability of our approach, we also present the results on the pictures of real driving scenes taken by our own camera (Fig. 5). It can be observed that our model provides plausible visual result even with different camera calibration.

Conclusion

This paper proposes an unsupervised solution to the problem of depth estimation. We unify the intermediate disparity maps to the resolution as same as input image, improving the hole error in final depth maps. Our results are superior to the baseline of Godard's work and classic fully supervised depth estimation networks. Moreover, our model has great ability of generalization to scenes untrained.

In future research, to improve the estimation accuracy, we will try to add video frames into the network to make use of the temporal information between frames. On the other hand, in order to detect the target comprehensively, 3-D detection of specific targets (such as pedestrians) using scene depth is a challenging problem worthy of our research.

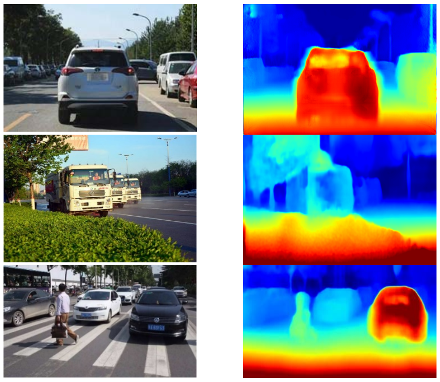

Figure 5. Results on outdoor driving dataset

ACKNOWLEDGMENT

This research has been supported by the National Natural Science Foundation of China (No.61673211, No.U1633105, No.61203170), the Foundation of the Graduate Innovation Center of Nanjing University of Aeronautics and Astronautics (kfjj20180702).

REFERENCES

[1] Bengler K, Dietmayer K, Farber B, et al. "Three Decades of Driver Assistance Systems: Review and Future Perspectives," IEEE Intelligent Transportation Systems Magazine, 2014, 6(4):6-22.

[2] Žbontar, Jure, Lecun Y. "Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches." The Journal of Machine Learning Research, 2016, vol. 17, pp. 2287-2318.

[3] Hirschmuller H. "Accurate and efficient stereo processing by semi-global matching and mutual information," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 2005, vol. 2, pp. 807-814

[4] Koenderink J J, van Doom A J. "Affine structure from motion," Journal of the Optical Society of America A Optics & Image Science, 1991,vol 8, pp. 377-385.

[5] Tang C , Hou C , Song Z. "Depth recovery and refinement from a single image using defocus cues," Journal of Modern Optics, 2015, vol 62, pp. 441-448.

[6] Prados E, Faugeras O. "Shape From Shading," Mathematical Models in Computer Vision the Handbook, 2009, vol 21, pp. 375-388.

[7] He Tong Neng, You Jia Geng, Chen De Fu. "Monocular image depth estimation based DenseNet," Computer Measurement & Control, 2019, vol 27, pp. 233-236.

[8] Bi T T, Liu Y, Weng D D, et al. "Survey on Supervised Learning Based Depth Estimation from a Single Image," Journal of Computer-Aided Design & Computer Graphics, 2018, vol 30, pp. 3-13.

[9] Eigen D, Puhrsch C, Fergus R. "Depth map prediction from a single image using a multi-scale deep network," Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT press, 2014, pp. 2366-2374.

[10] Zhou T, Brown M, Suavely N, et al. "Unsupervised learning of depth and ego-motion from video," IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2017, pp. 6612-6619.

[11] Xie J, Girshick R, Farhadi A. "Deep3D: Fully Automatic 2D-to-3D Video Conversion with Deep Convolutional Neural Networks," European Conference on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 842-857.

[12] Garg R, Bg V K, Carneiro G, et al. "Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue," European Conference on Computer Vision. Amsterdam, The Netherlands, 2016, pp. 740-756.

[13] Favaro P, Soatto S. "A Geometric Approach to Shape from Defocus," IEEE Transactions on Pattern Analysis & Machine Intelligence. 2005, vol 27, pp. 406-417.

[14] Cl´ement Godard, Oisin Mac Aodha, and Gabriel J Brostow. "Unsupervised Monocular Depth Estimation with Left-Right Consistency," IEEE Conference on Computer Vision and Pattern Recognition. Sep 2016, pp. 6602-6611, doi:10.1109/cvpr.2017.699.

[15] M.Jaderberg, K.Simonyan, A.Zisserman, and K.Kavukcuoglu. "Spatial transformer networks," ArXiv: 1506.02025, Mon 2015.

[16] Godard, Clément, Mac Aodha O , Firman M. "Digging Into Self-Supervised Monocular Depth Estimation," ArXiv: 1806.01260, Jun 2018.

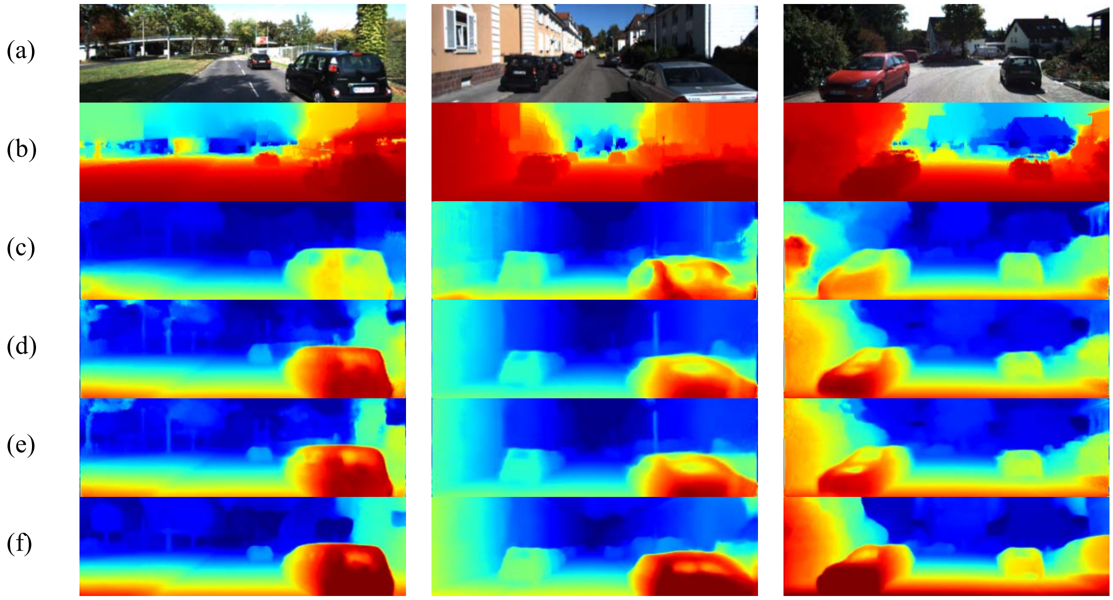

Figure 4. Comparisons of results using different approaches to estimate depth on KITTI dataset. (a) raw input data; (b) ground truth; (c) results of Eigen et al.'s method; (d) results of Zhou et al.'s method; (e) results of Godard et al; (f) results of our approach with multiscale unification