Perceptual losses for self-supervised depth estimation

Xiaoyu Liu a, Huachen Gao b, *, Xiaowen Ma c

School of Mechanical, Electrical& Information Engineering, Shandong University, Weihai, Weihai, China.

a 201800800425@mail.sdu.edu.cn
*, b Corresponding author e-mail:201800800405@mail.sdu.edu.cn,
c 201800800431@mail.sdu.edu.cn

Abstract. Convolution neural network has shown excellent results in stereo and monocular disparity estimation, while most of the existing methods convert the image depth prediction problem into the image reconstruction problem, and calculate the depth of each pixel through the disparity between the generated left and right images. However, in the reconstruction task, the loss is still calculated at the pixel level when comparing the reconstructed picture with the original picture, which will greatly affect the estimation of picture depth due to the problems of illumination and occlusion. Therefore, when calculating the loss of image reconstruction, it is very important to compare the higher-level features extracted from the reconstructed image with the original image. In this paper, based on the existing methods, we have innovated the loss function and introduced perceptual loss, i.e., we use feedforward neural network to extract features to further evaluate the reconstructed image, to make the reconstruction loss of baseline more accurate and improve the accuracy and robustness of the depth prediction model. To compare the improved effect, we performed extensive experiments on KITTI driving data by the improved model set, and the experimental index obtains better performance than the original baseline model.

Keywords: Perceptual loss, depth estimation, self-supervised.

1. Introduction

Depth estimation is a classic problem in the field of computer vision, which can be applied to robot navigation, augmented reality, 3D reconstruction, automatic driving, and other fields. Most of the works that were widely accepted about depth estimation are based on the transformation from two-dimensional RGB image to RGB-D image. Some methods include geometric image methods such as multi-view solid geometry (MVS), structure recovery in motion (SFM), and shape recovery from shadow (SFS) [1,2,3], as well as machine learning methods such as Markov random field (MRF) and conditional random field (7) [4,5], but these methods have some limitations such as considerable artificial assumptions and a large amount of computation. In order to solve these limitations, the application of the depth learning method in monocular image depth estimation has been explored.

To solve the difficulty and high cost of data acquisition with deep labels in deep learning methods, the number of methods that use unsupervised training for depth estimation has increased dramatically
in recent years. Godard et al. innovatively propose an unsupervised monocular depth estimation network[6] with left-right consistency which we call the MonoDepth. In this paper, the MonoDepth is used as the benchmark of our research. Their MonoDepth can estimate the disparity maps of the left view and the right view, and unsupervised training is carried out combined with luminosity loss, disparity smoothness loss, and left-right consistency loss. Although their consistency loss can greatly improve the performance of the network, there are still limitations, for example, the pixel-by-pixel comparison method used in photometric loss does not capture the perceptual difference between the reconstructed image and the original image. For example, consider two identical images offset by one pixel from each other. Although they are similar in perception, they are quite different in terms of pixel loss. At the same time, the pixel-by-pixel comparison method is greatly affected by uneven illumination and occlusion, and they are similar in perception. Therefore, when calculating the loss of image reconstruction, it is of great vital to compare the high-order features extracted from the reconstructed image with the original image.

Recent research [7] shows that high-quality images can be generated by defining and optimizing perceptual loss functions, which provide high-level features based on the use of pre-trained networks. Therefore, this paper makes an improvement on the Monocular depth estimation network. We train feed-forward prediction networks for depth estimation tasks. When training our networks, we use not only pixel-by-pixel loss functions but also perceptual loss functions that depend on high-level features from the pretrained networks. Our results are similar to the Monodepth [6] in speed, but the quality and the output of object function are improved. An example result of our algorithm is shown in Figure 1.

![Figure 1. Training results of our model on KITTI2015 stereo, from left to right: input picture, real disparity, model predicted disparity](image)

2. Related Work

2.1. Supervised Disparity Estimation

For monocular image depth estimation tasks, most early methods applied the supervised regression model, which means the model training data contains its corresponding depth label.

Eigen et al. [8] use two-scale neural networks to estimate the depth of a single picture: coarse-scale networks predict the global depth of the picture, and fine-scale networks optimize local details. The network model of Eigen et al. [9] uses a deeper basic network VGG Net, which includes three scales of networks. Scale1 network makes a rough estimation of the whole image, and then scale2 and scale3 networks are used to optimize the details of the global prediction. Liu et al. [10] combine depth convolution neural network with continuous conditional random field to propose depth convolution...
neural field to estimate depth from a single image. Liu et al. [11] improved on the overall framework improved by Trigueiros et al. [12], encoding super-pixel information into neural networks to improve computational efficiency. Li et al. [13] proposed a new multi-scale depth estimation method. Firstly, the depth neural network is used to regress the depth of the super-pixel scale, then the multi-layer conditional random field is used for processing, and the depth of the super-pixel scale and pixel scale is combined for optimization. Laina et al. [14] proposed a full convolution network architecture based on residual learning for monocular depth estimation. The network structure is deeper and does not need post-processing.

2.2. Unsupervised Disparity Estimation

The supervised learning method requires each RGB image to have its corresponding depth tag, but it is difficult and costly to obtain the depth tag. Therefore, recently a large number of algorithms are based on the unsupervised model, and their basic idea is to solve the depth by combining left and right views with epipolar geometry and an automatic encoder.

Garg et al. [15] use stereo images to realize unsupervised monocular depth estimation without depth tags. This work is similar to that of the automatic encoder. The training requires the stereo image pair composed of the original image and the target image. First, the encoder is used to predict the depth map of the original image, and then the decoder is used to reconstruct the original image by combining the target image and the predicted depth map, and the reconstructed image is compared with the original image to calculate the loss. Godard et al. [6] further improved the above methods, using left-right consistency to enhance unsupervised depth prediction. The disparity map is generated by epipolar geometry and the left-right consistency is used to optimize the performance and improve the robustness. Kuznietsov et al. [16] combine the supervised learning method labeled with sparse depth maps with the unsupervised learning method to further improve performance. The loss function of supervised learning promotes the depth map predicted by the network to be consistent with the depth map obtained by the sensor; The loss function of unsupervised learning promotes the consistency between the reconstructed image and the original image.

3. Method

This section will introduce our improvement on the depth estimation network for a single image. Our depth estimation system consists of two parts: A depth estimation network \( d_W \) and a loss network \( \phi \). The depth estimation network is a residual convolution neural network parameterized by weights \( W \), Specifically, we take Monodepth[6] as the basis of our work, which estimates the disparity maps of the left view and the right view, and combines luminosity loss, disparity smoothness loss and left-right disparity consistency loss for unsupervised training. The loss network based on the loss function of Monodepth [6] uses the method of Justin et al. [7] to extract high-level features from the original image \( I \) and the reconstructed image \( \hat{I} \) to calculate the perceptual loss. The perceptual loss is based on the pre-trained convolutional neural network VGG16 [17]. After four times of forward propagation, high-order features are extracted at a specific layer, in which the pre-trained network does not participate in the back propagation parameter update.
3.1. Depth Estimation Network Architecture

The purpose of the depth estimation network is to train a function that can predict the scene depth of an input image per pixel. Referring to Godard et al. [6], we convert the depth estimation problem to an image reconstruction problem during training. If we train a function to reconstruct an image observed at one position from an image observed at another position, then we can obtain the 3D shape of the scene and calculate the depth of each pixel through the parallax relationship between the two images. The specific network structure is depicted in Figure 3.

As for image reconstruction, the traditional method is that after calibration and epipolar correction, the disparity map \( d' \) is calculated by changing the pixel position of the corresponding point through the sum of two images \( I' \) and \( I \), i.e., the left and right color images. Then calculate the left figure through parallax to reconstruct the right figure. Assuming that the right figure is reconstructed by the left figure and disparity map \( I'(d') \) is \( \hat{I}_b \) (similarly, the left figure can also be reconstructed as mentioned above). At the same time, the calculation of image depth can also be calculated by disparities. Assume that \( d \) is the predicted disparity after image correction. Giving a baseline distance \( b \) and camera focal length \( f \) makes it possible to get the depth by \( \hat{d}_b = \frac{bf}{d} \).

In this article, we need to consider how to predict the right figure from the left figure through the neural network (the right figure is the same). If the right figure can be predicted, the depth can be calculated according to the left and right disparity. Therefore, we adopted the method of speculating the disparity map so that the left map can generate the right map. The method is to predict two disparities...
simultaneously with only the left figure and to enhance the prediction effect with the principle of left-right consistency. Therefore, the network generates the predicted image through binocular reverse matching, which is a completely differentiable network. This model only needs a single left image as the input of the network to complete the remaining work and then complete the prediction task, while the right image is only used during training.

3.2. Perceptual Losses

The concept of perceptual loss was first proposed by Justin et al. [7] and applied to picture style conversion. The loss function of the image conversion problem is commonly constructed at the pixel level. In other words, they trained a feedforward convolution neural network and defined the loss function using the difference between the predicted picture and the original picture at the pixel level. However, this loss does not capture the perceptual difference between the output image and the ground-truth image. Justin et al. applied the perceptual loss function to replace the conventional pixel-level error, which is mainly reflected in the fact that the perceptual loss uses a pre-trained model and uses a forward propagation to obtain a high-order feature map, making the picture be generated more semantically similar to the target picture than the original method. At the same time, in the process of extracting high-order feature maps, the parameters of the convolution neural network are not updated by back propagation, but only the forward calculation of the network is carried out once, which keeps the same prediction speed as the pixel-level loss.

There is a problem of perceptual similarity in our image reconstruction, that is, the two images are the same in perception, but they have pixel offset or uneven illumination. If measured by the original pixel-by-pixel comparison method [6], the two images that are similar in themselves will be very different, which is inconsistent with reality. Therefore, we define a perceptual loss function with reference to the above methods to measure the semantic differences between images at a high level. The perceptual loss is calculated by using the pre-trained image classification network $\phi$. In all our experiments, $\phi$ is a 16-layer VGG network pre-trained on ImageNet data set [17].

3.3. Loss Function

Previous networks, such as DispNet [18] and Monodepth [6], define loss functions using the difference between the original picture and the generated picture at the pixel level. We think that the difference between the two images cannot be obtained from the high-order semantic level, and simple pixel comparison cannot achieve the ideal effect. Therefore, we innovatively use the perceptual loss function proposed by Justin et al. [7] and modify it as a part of the loss function to evaluate the reconstructed image in a higher feature semantics, making the constructed image more similar to the original image.

We define $H_s$ as the loss on each output scale, then the loss output of the whole network should be $H = \sum_{s=1}^{4} H_s$. At each scale level, the loss function of each layer consists of four aspects, reconstruction loss, disparity smoothness loss, left-right consistency loss, and perceptual loss. The loss function of each layer is calculated as follows.

$$H_s = \alpha_{ap}(H_{ap}^r + H_{ap}^l) + \alpha_{ds}(H_{ds}^r + H_{ds}^l) + \alpha_{ir}(H_{ir}^r + H_{ir}^l) + \alpha_p(H_p^r + H_p^l)$$

where $H_{ap}$ represents the similarity between the reconstructed picture and the original input picture, $H_{ds}$ reveals the difference of smoothness between the two pictures, $H_{ir}$ enforces the left-right consistency and $H_p$ stands for the perceptual loss. Each item contains indicators for the left and right figures, but only the left figure is sent to the convolution layer.

Next, we will show the loss items of each part. The following is only an example of the left figure (e.g. $H_{ap}^l$). The situation on the right and left are similar.

**Reconstruction Loss**: Reconstruction loss function represents the difference between the
reconstructed left image $P_L^0$, and the original left image $I_L$, and can be defined as L1 and SSIM losses (similar to [6], the weight value $\alpha$ is used to roughly balance the influence of the two items on the final result). The reconstruction loss function can be defined as follows:

$$H_{ap} = \frac{1}{N} \sum_{i,j} \alpha (1 - \text{SSIM}(I_y^{ij}, P_y^0)) + (1 - \alpha) \| I_y^{ij} - P_y^0 \|_1$$

**Disparity smoothness Loss**: is similar to [6], and smoothness loss can give less penalty on the image gradient, as follows:

$$H_{ds} = \frac{1}{N} \sum_{i,j} \left| \partial_x d_y^{ij} \right| \| \phi \|_1 + \left| \partial_y d_y^{ij} \right| \| \phi \|_1$$

**Left-right (LR) consistency Loss**: is similar to [6], and the left-right consistency loss term is also used in our model. The loss of left and right consistency makes the generated left and right image differences more coordinated. The specific equation is as follows:

$$H_{lr} = \frac{1}{N} \sum_{i,j} \| d_y^{ij} - d_{y+r+e}^{ij} \|_1$$

**Perceptual Loss**: We innovatively add a perceptual loss term to the loss function to compare the consistency of original images and reconstruction images from high-order semantics. For example, the occluded areas in the reconstructed images are usually represented by highly deformed areas. Perceptual loss is a very ideal tool to detect and penalize these deformations. Moreover, we have changed based on perceptual loss to represent the perceptual loss of the image. In this part, we use the feature map output by the four layers (relu1_2, relu2_2, relu3_3, relu3_4) of the pre-training model VGG16 [17] as the input of our perceptual loss term, and the formula is as follows

$$H_p = \frac{1}{N} \sum_{l=1}^{4} \left\| \phi^l(I_L) - \phi^l(P_L^0) \right\|$$

### 4. Results
To ensure the accuracy of model evaluation, it is not allowed to select data sets lacking stereo pair data, so we choose the popular KITTI2015 data set for evaluation. We compared the MonoDepth model with the model of perceptual loss to judge the performance of our method.

#### 4.1. Implementation details
We choose the same training parameter settings as [6], $\alpha_{ap} = 1$, $\alpha_{lr} = 1$ and scale the disparity smoothing term with the subsampling factor $r$ of each size image: $\alpha_{ds} = 0.1/r$, $\alpha_p = 0.1$.

The network implemented on Pytorch has about 31 million trainable parameters, and it takes 24 hours to train 200 epochs on the data set using RTX3090 GPU. The learning rate of the first 100 epochs is 0.001, the learning rate is halved in the last 100 epochs, the batch size is 16, and the Adam optimizer is used, where $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$.

**Post-processing**: To overcome stereo occlusion, this paper adopts a solution like [6]. In addition to calculating the depth map $d$ for the input image $I$, the depth map $d'$ is also calculated for the horizontal displacement image of $I$, and then $d'$ is horizontally shifted back to form $d''$. Although $d$ and $d''$ are basically aligned, the two disparity maps are combined here to produce the final result, which is pp.

#### 4.2. KITTI 2015
All our models are only trained on the KITTI data set, which consists of the selection of 29,000 pairs of stereo pairs in the KITTI data set [24] and contains 33 scenes. Up to 15 cars and 30 pedestrians in each image, as well as various degrees of occlusion and truncation, are suitable for our model.
Table 1. Comparison of evaluation parameters of different models. Monodepth is the model of the original [6]. The results prove that the model with increased perceptual loss is the best.

| Model      | Dataset | Abs Rel | Sq Rel | RMSE | RMSE log | A1 | A2 | A3 |
|------------|---------|---------|--------|------|----------|----|----|----|
| Monodepth  | K       | 0.1016  | 1.7372 | 5.581| 0.180    | 0.905 | 0.964 | 0.983 |
| Monodepth pp| K       | 0.1007  | 1.5723 | 5.494| 0.177    | 0.904 | 0.964 | 0.984 |
| Ours       | K       | 0.0920  | 1.2444 | 5.127| 0.166    | 0.909 | 0.969 | 0.987 |
| Ours pp    | K       | 0.0897  | 1.1497 | 4.932| 0.162    | 0.911 | 0.969 | 0.987 |

Lower is better

Figure 4. Comparison of training results of different models on KITTI2015, from left to right: input picture, Monodepth, Monodepth pp, Ours, Ours pp

As displayed in Table 1, our network is better than the Monodepth benchmark in all indexes. Although a forward propagation of the VGG network affects our prediction speed making it slower than that of Monodepth, the quality and object function output has been improved well. Among them, Sq Rel and RMSE are obviously reduced, which indicates that the right view generated by our model from the left view is more similar to the real right view, and the image reconstruction effect is good. As shown in Figure 4, the depth map we generated is clearer and more accurate.

5. Conclusion
In this paper, we show the use of an improved unsupervised depth neural network to estimate the depth of a single picture. In this paper, MonoDepth model is our baseline, and binocular stereo image data which is easy to obtain is still used as training samples to replace real depth image data which is difficult and costly to obtain. Based on the original loss function, we innovatively incorporate the perceptual loss term. The use of this loss term can make the original image and the generated image more similar at a higher semantic level, not just at the pixel level. We take the feature maps of the pre-trained feedforward convolution neural network output original image and the reconstructed image at different levels as the input of the perceptual loss term, and the prediction time of a single forward propagation is close to that of the original model. Finally, the updated model is applied to KITTI2015 Stereo for testing, and the results show that our model is superior to baseline in all indicators. Therefore, our model has made progress in the original monocular depth estimation.

In our future work, we are going to adjust the model structure, such as considering the ambiguity of the image, increasing the weight for learning occluded or complex visual field areas, and considering reducing the number of channels in the self-encoder architecture to reduce parameters and lower model training costs.
References

[1] Zhang, Ruo, et al. "Shape-from-shading: a survey". IEEE transactions on pattern analysis and machine intelligence, vol. 21, no. 8, pp. 690-706, 1999

[2] Dellaert, Frank, et al. "Structure from motion without response". Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662). Vol. 2. IEEE, pp. 557-564, 2000.

[3] Dhand, Umesh R., and Jake K. Aggarwal. "Structure from stereo-a review". IEEE transactions on systems, man, and cybernetics, vol. 19, No. 6, pp. 1489-1510, 1989

[4] Saxena, Ashutosh, Sung H. Chung, and Andrew Y. Ng. "Learning Depth from Single Monocular Images", NIPS, Vol. 18, pp. 1-8, 2005

[5] Liu, Miaomiao, Mathieu Salzmann, and Xuming He. "Discrete-continuous depth estimation from a single image". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 716-723, 2014.

[6] Godard, Clément, Oisin Mac Aodha, and Gabriel J. Brostow. "Unsupervised monthly depth estimation with left-right consistency". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 270-279, 2017.

[7] Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution". European conference on computer vision. Springer, Cham, pp. 694-711, 2016.

[8] Eigen D, Puhrsch C and 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, Montreal: MIT Press, pp. 2366-2374, 2014.

[9] Eigen, David, and Rob Fergus. "Predicting depth, surface norms and semantic labels with a common multi-scale constitutional architecture.". Proceedings of the IEEE international conference on computer vision, pp. 2650-2658, 2015.

[10] Liu, Fayao, Chunhua Shen, and Guosheng Lin. "Deep constitutional neural fields for depth estimation from a single image". Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5162-5170, 2015.

[11] Liu, Fayao, et al. "Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 10, pp. 2024-2039, 2016.

[12] Trigueiros, Paulo, Fernando Ribeiro, and Luis Paulo Reis. "A comparison of machine learning algorithms applied to hand gesture recognition". 7th Iberian Conference on Information Systems and Technologies (CISTI 2012). IEEE, pp. 1-6, 2012.

[13] Li, Bo, et al. "Depth and surface normal estimation from monetary images using regression on deep features and hierarchical crfs". Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1119-1127, 2015.

[14] Laina, Iro, et al. "Deep Depth Prediction with Fully Convolutional Residual Networks". 2016 Fourth International Conference on 3D Vision (3DV). IEEE, pp. 239-248, 2016.

[15] Garg, Ravi, et al. "Unsupervised cnn for single view depth estimation: geometry to the rescue". European conference on computer vision. Springer, Cham, pp. 740-756, 2016.

[16] Kuznietsov, Yevhen, Jorg Stuckler, and Bastian Leibe. "Semi-supervised deep learning for monetary depth map prediction". Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6647-6655, 2017.

[17] K. Simonyan and A. Zisserman, "Very deep constitutional networks for large-scale image recognition networks for large-scale image recognition". In International Conference on Learning Representatives (ICLR), 2015.

[18] Mayer, Nikolaus, et al. "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation." Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4040-4048, 2016.