SEMI-SUPERVISED DEPTH ESTIMATION FROM A SINGLE IMAGE BASED ON CONFIDENCE LEARNING

Hu Tian, Fei Li

Fujitsu Research & Development Center Co., Ltd., Beijing, China

ABSTRACT

Recent advances in deep convolutional neural network have lead to significant improvement over depth estimation from a single image. However, training such networks usually needs a large amount of high-quality labeled training data which is difficult to collect. To alleviate it, a semi-supervised method based on confidence learning is proposed to improve the results of depth estimation with additional unlabeled data. We utilize the confidence map generated by a confidence network to predict the trusty regions of depth estimations on unlabeled data. These depth estimations are taken as pseudo ground truth and used to train the depth network together with labeled data. Experiments on NYU Depth dataset V2 show our proposed semi-supervised method outperforms fully-supervised method.

Index Terms— Depth estimation, confidence learning, semi-supervised, self-training

1. INTRODUCTION

Depth estimation from a single image is the task to assign a depth to each pixel in the image. It is essential to a wide range of applications, such as scene understanding and 3D reconstruction. Early works [1] [2] [3] for the task of depth estimation from a single image usually learn depth based on handcrafted features together with graphic models. This task remains challenging because of scene appearance variants, occlusion, and the lack of context understanding which are hard to be represented by these features.

Recently, Convolutional Neural Network (CNN) based methods have achieved astonishing performance and greatly improved the accuracy of depth estimation [4][5][6][7][8][9]. However, CNN-based methods require an enormous amount of training data. Though some consumer-level cameras such as Kinect can be used to capture depth maps, it also costs considerable expense and time.

To ease the effort of acquiring high-quality depth, we propose a semi-supervised method for the task of depth estimation from a single image based on confidence learning. Our proposed framework consists of two networks, i.e., depth network and confidence network. The depth network takes a single image as input and outputs a depth map which is a typical network for depth estimation. The aim of the confidence network is to predict a confidence map for the depth map estimated by the depth network. The confidence map indicates which regions in the estimated depth map are close to the ground truth. Then we utilize the confidence maps as the supervisory signal to guide the training of depth network using unlabeled images in a self-training manner. This idea is inspired by the methods in semi-supervised classification [10] where the class of unlabeled data with highest probability is taken as pseudo true label. However, depth estimation is a problem of regression instead of classification which means we cannot take the probability as confidence. Therefore, we propose to learn this kind of confidence for depth estimation by a neural network. Here it is assumed that confidence learning is a relatively easier task than depth estimation.

At the beginning, we jointly train the two networks in a supervised way. After the confidence network is trained well enough, the estimated depths for unlabeled images with high confidences are taken as ground truth depths and are used to further train the depth network together with labeled data. By adopting the proposed network, we show the depth estimation accuracy can be improved by adding unlabeled images.

2. RELATED WORK

Depth estimation Recent state-of-the-art methods for depth estimation from a single image are based on the rapid development of CNN. Eigen *et al.* [4] propose a multi-scale CNN for learning depth. Built upon this network, they [5] further develop a more general network with a sequence of three scales to refine predictions to higher resolution. Laina *et al.* [6] propose a more powerful single-scale fully convolutional residual network. The paper [7] presents a regression-classification cascaded network which can improve depth estimation results. To enforce spatial consistency of depth maps, depth gradient and graphical models are commonly used tools. A two-streamed CNN model is used in [8] to predict depth and depth gradient, which are then fused together into an accurate depth map. Xu *et al.* [9] combine the multi-scale CNN and continuous conditional random fields together and realize joint training by a novel mean-field updates. These methods are trained in a fully supervised way, however, pixel-level annotations are usually expensive and

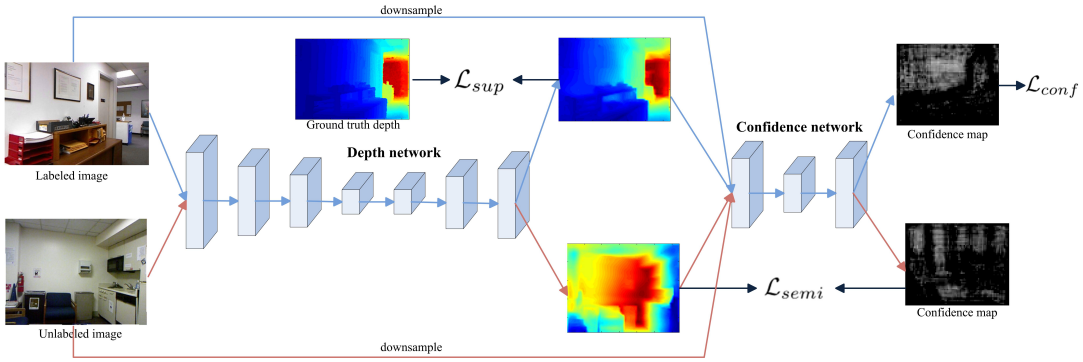

Fig. 1. Overview of the proposed framework for semi-supervised depth estimation based on confidence learning.

difficult to collect.

Semi-supervised learning Some methods perform unsupervised or semi-supervised depth estimation using stereo images at training [11] [12]. In this case, the left-right consistency is used as supervisory signal. But stereo images are also difficult to be collected. In the problem of semi-supervised classification, Lee [10] utilizes the maximum classification output as pseudo label for each unlabeled sample while this idea cannot be used for regression. In the task of semantic segmentation, Hung *et al.* [13] utilize the spatial probability map generated by a discriminator network as the supervisory signal to guide the cross-entropy loss for unlabeled data. However, this kind of probability map is difficult to represent the real confidences for estimated labels.

3. ALGORITHM OVERVIEW

Fig. 1 shows the overview of the proposed algorithm. Our framework is composed of two networks: the depth network and the confidence network. The former can be any network designed for depth estimation, e.g., Multi-scale CNN [5], ResNet-UpProj [6]. Given an input image with dimension $H \times W \times 3$, the depth network outputs the depth map of size $\frac{1}{4}H \times \frac{1}{4}W \times 1$ (small size due to memory concern).

Our confidence network is a fully convolutional neural network with Conv-Deconv architecture, which takes the image and estimated depth map as input and then outputs spatial confidence map (between 0 and 1) with size of $\frac{1}{4}H \times \frac{1}{4}W \times 1$. Each pixel p of the confidence map represents whether the estimated depth is accurate or not. For example, if the confidence value for pixel p is close to 1, it means the estimated depth for pixel p is likely to be close to the ground truth depth.

During the training process, we use both the labeled and unlabeled images under the semi-supervised setting. When using the labeled data, the depth network is supervised by the standard distance loss with ground truth depth. We represent the ground truth confidence as the similarity between estimated depth and ground truth depth. Then the confidence network is supervised by the standard distance loss with this ground truth confidence. Note that we train the confidence network only with labeled data.

For unlabeled data, we train the depth network with the proposed semi-supervised method. After obtaining initial estimated depth map of unlabeled image from the depth network, we then obtain a confidence map by passing the estimated depth map through the confidence network. We in turn treat this confidence map as the supervisory signal using a self-training scheme to train the depth network in the next epoch with a masked distance loss. The intuition is that this confidence map indicates the local quality of the estimated depth map, so that the depth network knows which regions to trust for unlabeled data during training.

4. SEMI-SUPERVISED TRAINING WITH CONFIDENCE LEARNING

In this section, we address the detailed learning scheme of the depth and confidence networks, as well as the network architecture.

Given an image $\mathbf{I}_n$ with size of $H \times W \times 3$, we denote the estimated depth map by the depth network as $\mathbf{E}_n$ which has the size of $\frac{1}{4}H \times \frac{1}{4}W \times 1$. For our fully convolutional confidence network, it takes the down-sampled image $\mathbf{I}_n$ and the estimated depth map $\mathbf{E}_n$ with the size of $\frac{1}{4}H \times \frac{1}{4}W \times 4$ as input and outputs a confidence map $\mathbf{C}_n$ with the size of $\frac{1}{4}H \times \frac{1}{4}W \times 1$.

4.1. Confidence Network Training

4.1.1. Training target for confidence network

We expect the confidence generated by the confidence network could represent the accuracy of depth estimated by the depth network, i.e., high confidence corresponds to high depth accuracy. To achieve this training target, we first define the ground truth confidence for a pixel p according to the relative error between estimated depth and ground truth

depth:

$$\mathbf{Y}_n(p) = \exp\left(\frac{-\alpha|\mathbf{E}_n(p) - \mathbf{D}_n(p)|}{\mathbf{D}_n(p)}\right) \quad (1)$$

where $\mathbf{D}_n(p)$ is the ground truth depth of pixel p and α is a constant. $\mathbf{Y}_n(p) = 1$ if the estimated depth is equal to the ground truth depth. Then the corresponding loss function reads as:

$$\mathcal{L}_{conf} = \sum_n \sum_p |\mathbf{C}_n(p) - \mathbf{Y}_n(p)| \quad (2)$$

By training the confidence network, it can predict a confidence map for the depth map estimated by depth network. If the confidences of some pixels are close to 1, it means the corresponding estimated depths have great probability to be close to the ground truth depths. We can select the depths with high confidences as pseudo ground truth depths for unlabeled images. Note that the ground truth confidence map $\mathbf{Y}_n$ is computed according to the depth map estimated by the depth estimation, so it will be updated at training. This means our confidence network is just learned for the depth network. Along with the improvement of the depth network at training, the confidence network is trained to predict accurate confidence for current depth estimation.

4.1.2. Confidence network architecture

For the confidence network, we adopt a Conv-Deconv architecture. The first three convolutional layers have $\{128, 256, 512\}$ channels with kernel size $\{3, 4, 4\}$ and stride size $\{1, 2, 2\}$ respectively. The following two deconvolutional layers have $\{512, 256\}$ channels with kernel size $\{4, 4\}$ and stride size $\{2, 2\}$ respectively. All these layers are followed by Leaky-ReLU. The last convolutional layer has one channel and is activated by sigmoid function to produce a confidence map. This architecture for the confidence network is simple than most of the networks for depth estimation.

4.2. Depth Network Training

4.2.1. Training target for depth network

We propose to train the depth network via minimizing the supervised and semi-supervised loss function:

$$\mathcal{L}_{dep} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{semi} \quad (3)$$

where $\mathcal{L}_{sup}$ and $\mathcal{L}_{semi}$ denote the supervised depth loss and semi-supervised depth loss, respectively. λ is a constant for balancing the two losses.

4.2.2. Training with labeled data

We first consider the scenario of using labeled data. Give an input image $\mathbf{I}_n$ and its ground truth depth map $\mathbf{D}_n$, the supervised depth loss is defined as the distance between estimated depth and ground truth depth:

$$\mathcal{L}_{sup} = \sum_n \sum_p |\mathbf{E}_n(p) - \mathbf{D}_n(p)| \quad (4)$$

Table 1. The effectiveness of confidence learning. The mean relative depth error is computed for different confidence thresholds on unlabeled data.

Labeled Data Amount	T_{semi}	rel
1/2	0.9	0.14
1/2	0	0.19
1/4	0.9	0.20
1/4	0	0.24

4.2.3. Training with unlabeled data

Now we consider the semi-supervised training with unlabeled data. For unlabeled data, we cannot apply $\mathcal{L}_{sup}$ since there is no ground truth depth available. Thus we propose to utilize the trained confidence network by a self-training strategy. The main idea is that the trained confidence network can predict a confidence map $\mathbf{C}_m$ for an unlabeled image $\mathbf{I}_m$, and it can infer the regions where the estimated depth results are close to the ground truth. We then binarize this confidence map with a threshold to highlight the trusty region:

$$\mathbf{B}_m(p) = \begin{cases} 1, & \text{if } \mathbf{C}_m(p) \geq T_{semi}; \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

where $\mathbf{B}_m$ denotes the confidence mask and T_{semi} is a threshold to control the sensitivity of self-training process. We represent the current estimated depth for $\mathbf{I}_m$ as $\hat{\mathbf{D}}_m$. Then the self-training ground truth is denoted as the masked estimation results $\{\hat{\mathbf{D}}_m, \mathbf{B}_m\}$ which will be used in the next training epoch. The resulting semi-supervised loss is defined by:

$$\mathcal{L}_{semi} = \sum_m \sum_p |\mathbf{E}_m(p) - \hat{\mathbf{D}}_m(p)| \cdot \mathbf{B}_m(p) \quad (6)$$

For minimizing Eq. 6 we treat $\{\hat{\mathbf{D}}_m, \mathbf{B}_m\}$ as constant, thus Eq. 6 is just a masked distance loss. Note that the masked estimation results $\{\hat{\mathbf{D}}_m, \mathbf{B}_m\}$ are updated for unlabeled image $\mathbf{I}_m$ at every training epoch. It means different trusty regions for unlabeled images will be used to minimizing the semi-supervised loss. This will help the depth network to avoid overfitting caused by noisy masked depth estimations for unlabeled data.

4.2.4. Depth network architecture

Almost all current CNN architectures involve a bottom-up pathway, which computes a feature hierarchy consisting of feature maps at several scales with a scaling step of 2, such as the classic VGG [14] and ResNet [15]. In our paper, we adopt

Table 2. Comparisons between fully supervised method and semi-supervised method on NYU Depth V2 test set.

Labeled Data Amount	λ	Error(lower is better)			Accuracy(higher is better)		
		rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
1/2	0	0.210	0.085	0.714	0.670	0.913	0.975
1/2	0.1	0.204	0.083	0.704	0.686	0.917	0.977
1/4	0	0.238	0.100	0.803	0.591	0.870	0.965
1/4	0.1	0.235	0.095	0.802	0.615	0.883	0.966

ResNet-50 model pre-trained on the ImageNet dataset as our base model. Because the output is a high resolution map in depth prediction, some forms of top-down pathway with upsampling are often required in order to obtain a larger depth map. Here we explore the effective architecture proposed in [16] for object segments, which uses a top-down architecture with lateral connection to build an in-network feature pyramid and makes prediction on the finest level. We extend it to the task of depth estimation. In the top-down architecture, all the 3×3 and 1×1 convolutional layers have 256 channels except that the last 3×3 convolutional layer has 1 channel to output a depth map.

5. EXPERIMENTS

5.1. Implementation Details

We implement our model using the MXNet framework [17]. The depth network is trained using mini-batch SGD where the momentum is 0.9 and the weight decay is 0.0005. The initial learning rate is set to 0.001 and decreased by a polynomial decay as mentioned in [18]. For training the confidence network, we adopt Adam optimizer with the learning rate as 0.0002. For parameter setting, we set α to 5, the confidence threshold T_{semi} to 0.9 and the weight λ to 0.1.

The depth network and confidence network are trained jointly for totally 100 epoches. To prevent the depth network suffering from initial noisy depth and confidence, we empirically start the semi-supervised learning after training for 40 epoches with labeled data. Note that confidence masks for unlabeled data are updated at each training epoch and used for semi-supervised learning at the next training epoch.

5.2. Evaluation Dataset and Metrics

To demonstrate the effectiveness of our proposed semi-supervised learning method, we train our model on publicly available NYU Depth V2 dataset [19]. This dataset contains 795 color-depth pairs for training and 654 for testing. The RGB images are downsampled to 320×240 to form the inputs of the depth network. Following [4], we augment the training data by random scaling, flips, color and translation which produces totally 7155 training image and depth map pairs.

We evaluate our predictions using the same measures as previous works [6] [8]: mean relative error (rel), root mean squared error (rms), mean log10 error (log10) and accuracy with threshold ($\delta < \left[1.25, 1.25^2, 1.25^3\right]$).

5.3. Results on NYU Depth V2 Dataset

To illustrate the effect of confidence learning, we examine the relationship between learned confidence and the accuracy of estimated depth. We firstly select the depth estimations for all unlabeled images with different confidence thresholds T_{semi} where we set it to 0.9 and 0 respectively. Then we compute the relative depth error for these depth estimations using ground truth. The results are listed in Table 1. $T_{semi} = 0$ means all depth estimations are selected. Obviously, the depth estimations with high confidence have smaller depth errors. This manifests we can select the trusty depth estimations for unlabeled images according to the confidence network. Taken them as weak supervised signal, they can help the depth network remember which regions are estimated accurately and thus improve the estimation results in future training.

To validate our proposed semi-supervised learning scheme, we randomly sample 1/2, 1/4 images as labeled data and use the rest of training images as unlabeled data. Table 2 shows the evaluation results with ($\lambda = 0.1$) and without ($\lambda = 0$) semi-supervised learning on NYU Depth V2 dataset. We can see that the semi-supervised loss brings consistent performance improvement on all six metrics over different amounts of training data. This verifies the effect of our proposed semi-supervised method for depth estimation.

6. CONCLUSIONS

In this paper, we propose a semi-supervised method for depth estimation based on confidence learning. We train a confidence network to enhance the depth network with both labeled and unlabeled data. For unlabeled images, the confidence maps generated by the confidence network act as the self-training signal to refine the depth network. Experiments on NYU Depth V2 dataset is performed to validate the effectiveness of the proposed method. In the future, we will study the performance as a large amount of unlabeled data are provided.

7. REFERENCES

[1] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng, "Learning depth from single monocular images," in *Advances in neural information processing systems*, 2006, pp. 1161–1168.

[2] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng, "3-D depth reconstruction from a single still image," *International Journal of Computer Vision*, vol. 76, no. 1, pp. 53–69, 2008.

[3] Beyang Liu, Stephen Gould, and Daphne Koller, "Single image depth estimation from predicted semantic labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1253–1260.

[4] David Eigen, Christian Puhrsch, and Rob Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

[5] David Eigen and Rob Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.

[6] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proceedings of the IEEE International Conference on 3D Vision*, 2016, pp. 239–248.

[7] Huan Fu, Mingming Gong, Chaohui Wang, and Dacheng Tao, "A compromise principle in deep monocular depth estimation," *arXiv preprint arXiv: 1708.08267*, 2017.

[8] Jun Li, Reinhard Klein, and Angela Yao, "A two-streamed network for estimating fine-scaled depth maps from single RGB images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 22–29.

[9] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe, "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[10] Dong-Hyun Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proceedings of ICML Workshop on Challenges in Representation Learning*, 2013.

[11] Ravi Garg, Kumar B. G Vijay, Gustavo Carneiro, and Ian Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 740–756.

[12] Clement Godard, Oisin Mac Aodha, and Gabriel J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6602–6611.

[13] Wei Chih Hung, Yi Hsuan Tsai, Yan Ting Liou, Yen Yu Lin, and Ming Hsuan Yang, "Adversarial learning for semi-supervised semantic segmentation," *arXiv preprint arXiv: 1802.07934*, 2018.

[14] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv: 1409.1556*, 2014.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[16] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár, "Learning to refine object segments," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 75–91.

[17] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang, "MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.

[18] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *arXiv preprint arXiv:1606.00915*, 2016.

[19] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from RGBD images," in *Proceedings of European Conference on Computer Vision*, 2012, pp. 746–760.