Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture (arXiv 2014)

Seminar *Recent Trends in 3D Computer Vision*

Christoph Baur <c.baur@tum.de>
Supervisor: Benjamin Busam

Chair for Computer Aided Medical Procedures
Technische Universität München, Germany

Abstract. This paper presents a special CNN architecture which set the state-of-the-art of three different computer vision tasks at the same time, notably depth and surface normal prediction in monocular images as well as semantic labeling of RGB(-D) data. The CNN architecture operates at multiple scales, each of which also incorporates the output from the previous scale for progressive prediction refinements. This paper embeds the approach into the context of related work, presents the proposed architecture and shows selected results.

1 Introduction

Within the last decade, the large-scale availablity of low-cost depth sensors gave rise to novel methods for 3D scene understanding, which until then was restricted to stereo-view or motion approaches[21]. RGB-D datasets such as the NYUDepth[22] dataset were constructed and machine learning for physical geometry regression from single RGB images alone finally became possible thanks to the availablity of training data. Consequently, many methods for solving the challenging, deeply connected tasks of depth[11, 12, 3, 1, 15] and surface normals estimation[5, 18, 25, 4, 6] from monocular images have been proposed. Further, the availability of RGB-D data lead to novel approaches for semantic labeling[27, 10, 24, 13, 22, 2, 19, 8, 26, 9], which also incorporate depth information. While in particular the tasks of depth and normals prediction from RGB images are ill-posed problems and thus not completely accurate, such predictions can provide valuable information for other applications such as human pose estimation or robot navigation[1, 3, 22]. Recently, there has been one particular method which outperformed any of the previous contributions in all of the aforementioned tasks. The mentioned method is based on a single multi-scale CNN architecture which can easily be adapted to the different tasks with minor modifications. This report is an attempt to first give an overview of related work and briefly explains the concept of CNNs. In section 4, the proposed CNN architecture, the modifications, training and results are presented in greater detail.

2 Related work

A first attempt towards automatic depth prediction from single-view outdoor images was made by Hoiem et al. [11] with an automatic photo pop-up algorithm, in which they categorize parts of an image into three different geometric classes: ground, vertical and sky. "Cutting and folding" along the boundaries of the detected vertical regions yields the photo pop-ups. However, this approach is restricted to simple outdoor scenes and also not very accurate. Also in 2005, Saxena et al.[20] made a more sophisticated attempt for inferring depth from single monocular images, both indoor and outdoor. In their Make3D framework, they model the relationship between a variety of image features and depth maps directly as a Markov Random Field. In a later publication (2007), they improved their model and enhanced the input features. In 2012, Karsch et al.[12] proposed a method for automatic depth map recovery for indoor scenes based on non-parametric sampling. Their idea does not require any training. Instead, they rely on a database of RGB-D images where they choose samples from. A prediction is obtained by finding the best sample consensus based on minimizing an energy function. Ladickey et al.[15] argue that "there are mutual dependencies between the visual appearance of a semantic class and its geometric depth" and thus suggest to solve the problem of semantic labeling in RGB images and depth prediction jointly with a "semantic depth classifier". Baig et al.[1] proposed a method closely related to the method of Karsch et al. In contrast, Baig et al. do not rely on a dataset of complete RGB-D samples, but rather on two dictionaries containing compressed representations of images and depth. Based on a simple linear mapping between the dictionary-dependent representations of RGB and depth images, they can infer depth. Ladickey et al.[18] believe that the direct recovery of depth maps from monocular images is of limited use and suggest to rather predict surface normals. For this purpose, they propose a continuous boosting framework which regresses from both contextual features and segments to surface normals. In preceding work, Fouhey et al.[5] proposed a framework for surface normal estimation on indoor RGB-D data with help of learned 3D primitives that are both easy to recognize and carry useful 3D information. In [6], Fouhey et al. fuse these 3D primitives with a global optimization on a grid obtained from rays through the vanishing points of a scene in order to obtain more accurate surface normals. Meanwhile, the computer vision community has evolved a lot in the field of semantic labeling. While earlier approaches were based on heavy optimization problems or classifiers trained on handcrafted features, state-of-the-art results have been achieved with CNN architectures. For instance, both Couprie et al. and Farabet et al.[2, 4] proposed CNN architectures applied to the input at different scales, building a consensus on the separate predictions afterwards. Gupta et al. argue that raw depth features should be substituted by a special geocentric depth encoding, which together with a CNN would lead to better semantic labeling. Similar to Couprie et al., Eigen et al.[3] also proposed a multi-scale architecture in 2014, however for the task of depth prediction from monocular images. Their method consists of a network making global predictions whose output is fed into a second network specialized for making more local

predictions, in combination with a special scale-invariant loss-function. This work can be considered the predecessor of the outperforming architecture presented in this report.

3 Convolutional Neural Networks

Deep CNNs have first been introduced in 1998 by LeCun et al.[16] for document recognition and since then have set the state-of-the-art in many fields. Just recently they also set the state of the art in various computer vision tasks such as image classification[14], object detection[7] or stereo matching[28]. CNNs enhance the concept of neural networks by so called convolutional layers, and sometimes by pooling layers for subsampling. In contrast to ordinary neural networks, where each neuron is fully connected to all neurons of the previous layer, convolutional layers introduce a sparse, more local connectivity, which allows for deeper architectures and faster training. The term "convolutional" is coined by the fact that the learned weights of these sparse connections can be considered the coefficients of filters subject to convolution operations. This convolutional property renders CNNs particularly suited for computer vision tasks.

4 The Contribution

Eigen and Fergus propose a single, versatile, multi-scale CNN architecture which can be adapted to different tasks with minor modifications. A trained model starts off with a coarse, global prediction from the entire input image and then progressively makes local refinements of the predictions at successive scales.

4.1 Architecture

The proposed architecture (Fig. 4.1) consists of overall 3 different scale networks which progressively refine predictions based on the original input and the predictions from the previous scale. When feeding the models with an input image at the size of 320x240px, at the last scale an output at approximately half the input size is returned.

Scale 1 is an AlexNet[14] network which is used to compute multiple feature maps based on the entire image field of view. There are multiple convolution and pooling blocks which extract features and consecutively reduce the input size, as well as two fully connected (FC) layers at the end. The last FC layer outputs 64 feature vectors which are reshaped to the size of 19x14, which is approx. 1/16 of the input size.

Scale 2 is intended to do mid-level predictions. The original input image is first convolved and pooled with 32x9x9 filters, which leads to 32 feature maps with a size of 74x55. These feature maps are augmented with the up-sampled 64 feature maps from scale 1, leading to total 96x74x55 feature maps. The remaining layers of this scale are fully convolutional, such that the output is of the resolution

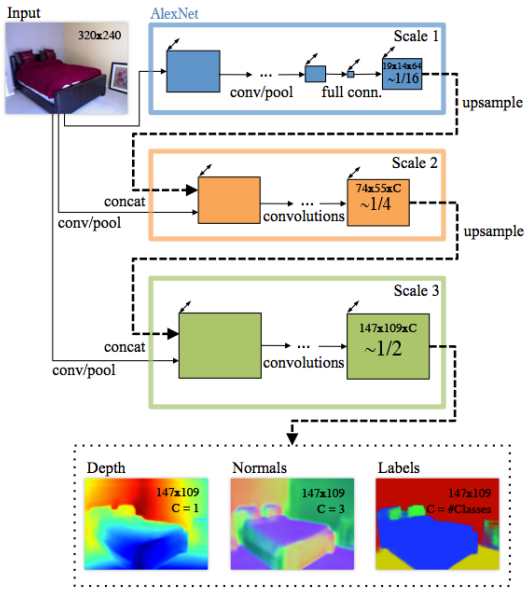

Fig. 1. Illustration of the multi-scale CNN architecture proposed by Eigen and Fergus

74x55xC. The number of output channels C of the prediction depends on the task and can be adjusted, of course.

Scale 3 acts similar to scale 2. The input image is again first convolved and pooled with 32x9x9 filters, but at a lower stride. This results in feature maps at the size of 147x109px, which are again augmented with the up-sampled output from scale 2. The resulting feature maps are subject to several more convolutions. The final output is of the size of 147x109xC, where C is the number of channels, respectively.

What renders this architecture "generic" is the fact that in order to adapt it to different tasks, simply the loss function has to be exchanged and a few parameters have to be tweaked.

4.2 Depth Prediction

The goal of the depth prediction task is to infer the absolute depth of each pixel in an RGB image. For this task, the authors set $C = 1$ and train the model with the following elementwise loss:

$$L_{depth}(D, D^*) = \frac{1}{n}\sum_i d_i^2 - \underbrace{\frac{1}{2n^2}(\sum_i d_i)^2}_{-\frac{1}{2n^2}\sum_i d_i^2 - \frac{1}{n^2}\sum_{i\neq j} d_i d_j} + \frac{1}{n}\sum_i [(\nabla_x d_i)^2 + (\nabla_y d_i)^2] \tag{1}$$

, where $d_i = log(\frac{D}{D^*})$ with D being the predicted and D^* being the ground truth depth map. The first term is the l2-error which is to be minimized. In order to see the impact of the second term more clearly, it has to be split up. The split yields a second order l2-error term which will dampen the actual l2-error. The second term of the split can be considered a scale consistency penalizer: it will impose a heavy penalization on the loss if d_i and d_j, i.e. two different relative pixel-errors being compared, have different signs, i.e. when one pixel is in front of its corresponding groundtruth pixel and the other one is behind. On the other hand, it awards the loss when both pixels being compared have the same sign, i.e. are consistent in scale. The third term of the loss enforces a spatial consistency and tries to preserve local structures by matching edges of predictions and groundtruth.

4.3 Surface Normals Prediction

The goal of the surface normals task is to predict a 3-component vector at each pixel of an image, i.e. its x,y and z components. For this task, the authors propose a simple loss function based on the dot produt of the predicted and the groundtruth normal vector:

$$L_{normals}(N, N^*) = -\frac{1}{n}\sum_i N_i \cdot N_i^* = -\frac{1}{n} N \cdot N^* \tag{2}$$

In fact, the elementwise loss compares the directions of the predicted to the corresponding groundtruth normal. Notably, for this task the authors set $C = 3$.

4.4 Semantic Labeling

For semantic labeling, i.e. predicting a class label for each pixel in an image, the authors make use of a standard softmax-classifier in conjunction with the cross-entropy loss. The latter plays nicely together with the softmax activation since it prevents vanishing gradients in sigmoid activations that originally would lead to a dramatic training slow down.

$$L_{semantic}(C, C^*) = -\frac{1}{n}\sum_i C_i^* \log(C_i) \tag{3}$$

$$C_i = \frac{\exp z_i}{\sum_c \exp z_{i,c}} \tag{4}$$

For the semantic labeling task, the authors make some changes to the architecture. At first, the input of scale 2 and 3 is augmented with groundtruth depth and normals for these experiments. However, Eigen and Fergus do not simply convolve all input channels together with a single set of filters, but apply a different set of 32x9x9 filters for each input type separately. This way, independence of the resulting feature maps is enforced which improved performance according to the authors. Further, they adjust the C parameter such that the number of output channels equals the number of class labels.

4.5 Training and Datasets

Eigen and Fergus conduct their experiments on two datasets. They run the experiments on the NYUDepth v2 dataset[22], consisting of heavily cluttered indoor scenes, for all tasks. Additionally, and only for the semantic labeling task, they also experiment with the SiftFlow dataset of land- and cityscapes[17]. In order to train a model for one of the tasks, they first extract training data from the respective dataset and perform the standard data augmentation, i.e. translation, scaling, in-plane rotations and more to increase the number of training samples. For all of the tasks, the convolutional layers of scale 1 are initialized from ImageNet trained weights to speed up convergence. The weights of all other layers are initialized randomly. Afterwards, they train the models in two phases. In phase 1, they attach the respective loss function to scale 2 and train both scale 1 and 2 jointly using SGD[14] from 5M samples. Recall that scale 1 yields 64 different feature maps rather than an actual prediction, hence the loss is attached to scale 2 right away. In phase 2, the trained weights and parameters of scale 1 and 2 are fixed, and the loss function is attached to scale 3. Then, they train the weights of scale 3, again on 5M samples. Notably, Eigen and Fergus do not train scale 3 using the entire input images as it is computationally heavy. Instead, they use random crops of size 74x55 of both the up sampled scale 2 output and original input data, which leads to a 3x training speed up.

4.6 Experiments

For all three tasks, Eigen and Fergus employ the NYUDepth dataset for training and testing. They train their models using the official train/test split[1] of the dataset and test on 654 images.

In their experiments on depth prediction, they originally compare the performance of their model to the methods from Karsch[12], Baig[1] and their own predecessor[3] by various metrics(Table 4.6). However, all of these methods have only one metric in common, which is the RMSE metric. Based on this metric, Eigen and Fergus show a significant relative improvement compared to their predecessor two-scale method of 14%, and compared to Baig et al. they even measure a relative improvement of 32%. This shows how adding a third scale to the network increases performance.

In their experiment on surface normal prediction, they compare to two different methods proposed by Fouhey et al.[5, 6], and to Ladickey et al.[18] in terms of the mean angle distance(mean of the cosine similarity), median angle distance(median cosine similarity) and the percentage of normals being less than 11.25% different from groundtruth(Table 2). In fact, they conduct different experiments with different groundtruth data. This is due to the fact that the NYUDepth dataset does not have groundtruth normals, instead these can be computed from depth data with different algorithms. The groundtruth for

[1] `http://cs.nyu.edu/~silberman/projects/indoor_scene_seg_sup.html`

Table 1. Comparison of depth prediction performance of the presented method vs other methods in terms of the RMSE

Metric	Karsch[12]	Baig[1]	Eigen[3]	This
RMSE	1.2	1.0	0.877	**0.753**

Table 2. Surface normal prediction performance comparison in terms of the mean and median angle distance as well as the percentage of normals that are within than 11.25% compared to groundtruth

Metric	Ladickey[18]	Fouhey[5]	Fouhey[6]	This
Mean Angle Distance	32.5	34.2	35.1	**23.1**
Median Angle Distance	22.3	30.0	19.2	**15.1**
% Within 11.25° Degrees	27.4	18.5	37.6	**39.4**

Table 3. Semantic labeling performance comparison in terms of the pixel accuracy for 4, 13 and 40 classes on the NYUDepth v2 dataset

Classes	Couprie[2]	Khan[13]	Gupta 13[9]	Gupta 14[8]	This
4	64.5	69.2	78.0	-	**80.6**
13	52.4	58.3	-	-	**70.5**
40	64.5	69.2	59.1	60.3	**62.9**

results in table 2 has been obtained according to [5]. However, Eigen and Fergus beat the other methods in all the mentioned metrics and on all different groundtruth datasets.

For semantic labeling, they perform experiments on two different datasets: the NYUDepth dataset of indoor RGB-D data, and the SiftFlow dataset of outdoor land- and cityscapes. On the NYUDepth dataset they train different models for predicting 4 classes, 13 classes and 40 classes(Table 3). In succession, they compare to many other methods in terms of the pixel accuracy, the per-class pixel accuracy, the Jaccard Index and the mean pixel-frequency weighted Jaccard Index. Again, the multi-scale CNN significantly beats all the other methods, except for the 40 class task. In the latter, their model is at least competitive to Gupta et al. 2014.

Eigen and Fergus also train models on the SiftFlow dataset, showing that the network can deal with different data. However, this is beyond the scope of this report. For the SiftFlow results as well as a visualization of predicted depth, normals and semantic labels, the interested reader is referred to the original paper.

In addition to these experiments, the authors also investigate how each scale contributes to the overall performance by training models consisting only of a

particular scale or a combination of scales. As a result, progressive improvements are experienced as more scales are added, which justifies the choice of the multi-scale architecture. Notably, for the depth and normals task, scale 1 makes the largest contribution. For the semantic labeling task it is scale 2, since groundtruth depth and normals are added at this scale which take precedence over the predicted features from scale 1. This circumstance is in fact examined in additional experiments, where the authors investigate the importance of adding groundtruth depth and normals to the network input. As a baseline, they do both training and prediction with a scale 2 only model on RGB data. When they add scale 1, they notice a small improvement. When they add groundtruth depth and normals to the scale 2 and 3 input, they notice a considerable improvement. Further, they also replace the groundtruth depth and normals with predicted depth normals and compare the performance. Doing so, they notice that the model performs much like the RGB-only two-scale model. This shows that scale 1 itself is capable of extracting enough depth and normals information on its own. Indeed, the coarse scale 1 network is very important for predictions from pure RGB data in all tasks.

5 ICCV 2015

In the meantime, Eigen and Fergus improved their architecture and submitted a revised version of their paper to ICCV 2015. They replaced the scale 1 AlexNet by the ImageNet-winning VGG-Net[23] architecture which has a larger receptive field size. With this modification, they, for instance, were able to improve the RMSE of the depth prediction task by 14% compared to their arXiv 2014 model. Furthermore, they also ran experiments on the PASCAL VOC dataset in order to be able to compare to other methods which had meanwhile been proposed for the three vision tasks. Finally, they were also able to successfully combine the depth and normals prediction task into a single network which shares scale 1 and allows for prediction of depth and normals in parallel, including a 1.6x training speed up.

6 Conclusion

In conclusion, Eigen and Fergus made a very valuable contribution to the computer vision community with a single, multi-scale CNN architecture that can easily be adapted to different tasks by formulating an appropriate loss function and adjusting a few parameters. They justify the choice of their architecture by setting the state-of-the-art for three different vision tasks. Notably, their method does not require any pre- or postprocessing, i.e. the models can be trained end-to-end, and it operates in realtime (30Hz) on a GPU. In future work they aim to use sparsely labeled data for faster training and suggest to also apply their method to other tasks such as instance labeling.

References

1. Mohammad Haris Baig, Vignesh Jagadeesh, Robinson Piramuthu, Arpit Bhardwaj, Wei Di, and Neel Sundaresan. Im2depth: Scalable exemplar based depth transfer. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 145–152. IEEE, 2014.
2. Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013.
3. 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*, pages 2366–2374, 2014.
4. Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Scene parsing with multiscale feature learning, purity trees, and optimal covers. *arXiv preprint arXiv:1202.2160*, 2012.
5. David F Fouhey, Arpan Gupta, and Martial Hebert. Data-driven 3d primitives for single image understanding. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3392–3399. IEEE, 2013.
6. David Ford Fouhey, Abhinav Gupta, and Martial Hebert. Unfolding an indoor origami world. In *Computer Vision–ECCV 2014*, pages 687–702. Springer, 2014.
7. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jagannath Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014.
8. Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Computer Vision–ECCV 2014*, pages 345–360. Springer, 2014.
9. Swastik Gupta, Pablo Arbelaez, and Jagannath Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 564–571. IEEE, 2013.
10. Alexander Hermans, Georgios Floros, and Bastian Leibe. Dense 3d semantic mapping of indoor scenes from rgb-d images. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 2631–2638. IEEE, 2014.
11. Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. *ACM Transactions on Graphics (TOG)*, 24(3):577–584, 2005.
12. Kevin Karsch, Ce Liu, and Sing Bing Kang. Depth extraction from video using non-parametric sampling. In *Computer Vision–ECCV 2012*, pages 775–788. Springer, 2012.
13. Salman Hameed Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Geometry driven semantic labeling of indoor scenes. In *Computer Vision–ECCV 2014*, pages 679–694. Springer, 2014.
14. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
15. Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 89–96. IEEE, 2014.
16. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

17. Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. Sift flow: Dense correspondence across different scenes. In *Computer Vision–ECCV 2008*, pages 28–42. Springer, 2008.
18. Bernhard Zeisl L'ubor Ladickỳ and Marc Pollefeys. Discriminatively trained dense surface normal estimation.
19. Andreas C Muller and Sven Behnke. Learning depth-sensitive conditional random fields for semantic segmentation of rgb-d images. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 6232–6237. IEEE, 2014.
20. Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems*, pages 1161–1168, 2005.
21. Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.
22. Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012*, pages 746–760. Springer, 2012.
23. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
24. Jörg Stückler, Benedikt Waldvogel, Hannes Schulz, and Sven Behnke. Dense real-time mapping of object-class semantics from rgb-d video. *Journal of Real-Time Image Processing*, pages 1–11, 2013.
25. Joseph Tighe and Svetlana Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3001–3008. IEEE, 2013.
26. Anran Wang, Jiwen Lu, Gang Wang, Jianfei Cai, and Tat-Jen Cham. Multi-modal unsupervised feature learning for rgb-d scene labeling. In *Computer Vision–ECCV 2014*, pages 453–467. Springer, 2014.
27. Xiaolong Wang, Liliang Zhang, Liang Lin, Zhujin Liang, and Wangmeng Zuo. Deep joint task learning for generic object extraction. In *Advances in Neural Information Processing Systems*, pages 523–531, 2014.
28. Jure Žbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. *arXiv preprint arXiv:1409.4326*, 2014.