DepthSegNet

Krishna Bhatu
116304764
Meng. Robotics
kbhatu@terpmail.umd.edu

Hrishikesh Tawade
116078092
Meng. Robotics
htawade@terpmail.umd.edu

Aditya Vaishampayan
116077354
Meng. Robotics
adityav@terpmail.umd.edu

Abstract

Our project addresses the two most important task in computer vision of depth estimation and semantic segmentation. Generally, both the tasks are addresses separately, but we have considered the idea of integrating both the task together in the same framework as they may benefit each other in improving the accuracy. We have integrated both the depth estimation network and the semantic segmentation network together into a single convolutional network. The main goal is to keep best features of both the networks and getting better results without compromising accuracy. Qualitative and quantitative experiments demonstrate that the performance of our methodology outperforms the state of the art on single-task approaches, while obtaining competitive results compared with other multi-task methods.

Introduction

Semantic and depth are intrinsically related and considering both of them together will aid in applications such as Robotics and autonomous navigation. In robotics performing tasks in highly interactive environments requires need of identifying objects and their correct distance from the sensors and the camera. Autonomous navigation applications need 3D reconstruction of scene and semantic information which ensures enough information, so that the agent can't carry out the required task autonomously. Also, RGB-D sensors are currently being used in many applications, most systems only provide RGB information. Therefore, addressing depth estimation and semantic segmentation under a unified framework is of special interest. Semantic depth maps provide represent the geometrical information of the scene and thus integrating the depth estimation and semantic segmentation into a single network is beneficial. Also because of the prior knowledge the feature extractors can also be trained in a better way. Hence, we have built a model which contains features that are extracted for both the tasks of segmentation and depth estimation. Also the main benefit of this architecture is that both semantic segmentation and depth maps are obtained from only a single image and thus it provides a solution.

Related work

A unified network containing local and global prediction where consistency between semantic segmentation and depth is learned, via a joint training process. An input image is given to the CNN and the depth map and the segmentation mask are jointly predicted. Then the image is further decomposed into local parts and the regions are trained on another CNN thus predicting the depth and segmentation for each individual egion. Thus with such two layered prediction I.e. the global and local predictions the problem is formulated into two layered conditional random field problem thus producing the final depth and segmentation map.

Another methodology proposes that initial level estimation for depth and semantic labels at pixel level via a unified network. At a later stage depth estimation is used for solving confusions between similar semantic categories thus obtaining a final segmentation map.

Another architecture called Multinet is proposed that performs segmentation, classification and depth estimation simultaneously. All the three tasks are incorporated into a unified decoder and encoder network. The outputs here are predicted in real time.

These work efforts were focused on improving the computational efficiency for real-time applications as autonomous driving.

Also another approach of Pixel-level Encoding and Depth layering was referred where an FCN was trained for estimating labels at pixel level.

Data

For our task we required labels for semantic segmentation along with the stereo images for depth estimation. Cityscape contains images from a diverse set of stereo video sequence recorded in street scene in 50 different cities and 30 classes. It contains 5000 high-quality pixel-wise annotations and 20000 weakly annotated images. Right and left stereo images are available for the same dataset. We can exploit the large weakly annotated data by pre training the network on high-quality labels and using the weights to initialize the network for weakly annotated learning.

Training losses

We define C_s at each output scale s, thus forming the total loss as the sum $C = \sum_{s=1}^{4} C_s$. The depth losses C_s are a combination of three main terms.

$$C_s = \alpha_{ap}(C_{ap}^l + C_{ap}^r) + \alpha_{ds}(C_{ds}^l + C_{ds}^r) + \alpha_{lr}(C_{lr}^l + C_{lr}^r)$$

Here C_{ap} encourages the reconstructed image to appear similar to the corresponding training input, C_{ds} enforces smoothness disparities, and C_{lr} prefers the predicted left and right disparities are constant.

Appearance matching loss:

$$C_{ap}^l = \frac{1}{N}\sum_{i,j} \alpha \frac{1-\text{SSIM}(I_{ij}^l, \tilde{I}_{ij}^l)}{2} + (1-\alpha)\left\|I_{ij}^l - \tilde{I}_{ij}^l\right\|$$

Disparity smoothness loss:
Disparities are preferred to be locally smooth with an L1 penalty on the gradients.

$$C_{ds}^l = \frac{1}{N}\sum_{i,j} \left|\partial_x d_{ij}^l\right| e^{-\left\|\partial_x I_{ij}^l\right\|} + \left|\partial_y d_{ij}^l\right| e^{-\left\|\partial_y I_{ij}^l\right\|}$$

Left right Disparity consistency Loss
So as to create accurate disparity maps, the network is trained to predict both left and right image disparities, even after being given only a left side image as input to the CNN. So as to ensure coherency, the left right disparity was introduced. This cost attempts to make the left-view disparity map be equal to the projected right-view disparity map.

$$C_{lr}^l = \frac{1}{N}\sum_{i,j} \left|d_{ij}^l - d_{ij+d_{ij}^l}^r\right|$$

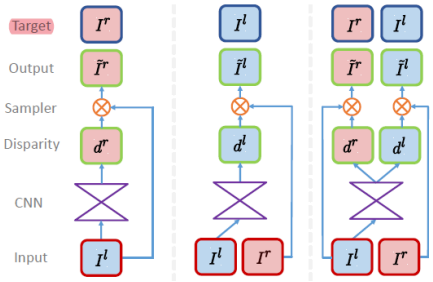


Fig: Disparity network

Cross entropy loss:

$$J = -\frac{1}{N}\left(\sum_{i=1}^{N} \mathbf{y_i} \cdot \log(\hat{\mathbf{y}}_\mathbf{i})\right)$$

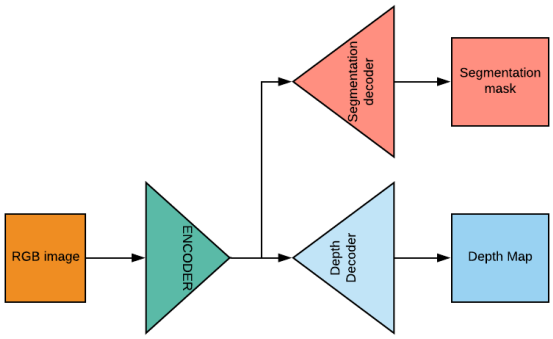
This loss generally measures classification performance with an output between 0 and 1. With the divergence of the predicted probability from true labels the loss increases.

Implementation details

The network which is implemented in TensorFlow contains the depth encoder and encoder which is initially trained on cityscape images with semantic segmentation labels. Thus, after training we get a decoder that Is suitable for semantic segmentation. Later we add the decoder obtained to the existing depth network. The network that was trained using cityscape images contains 31 million trainable parameters and takes on the order of 30 hours to train using a single Nvidia GTX 1080Ti GPU on a dataset of 20 thousand weakly annotated images for 50 epochs. ELU was used instead of RELUs for the non-linearities in the network.
Batch norm wasn't used as it didn't produce any significant difference in result. Later the segmentation decoder from the trained network was added to the depth network. And the whole network was trained again for 14 hours, 50 epochs on a NVIDIA TESLA GPU. The default parameters used were a batch size of 8, an Adam optimizer and a learning rate of 10^{-4}

which was kept constant during first 30 epochs and was halves after every 10 epochs.

Underlying architecture



ENCODER

layer	k	s	chs	In	Out	input
Conv1	7	2	3/32	1	2	Left
Conv1b	7	1	32/32	2	2	Conv1
Conv2	5	2	32/64	2	4	Conv1b
Conv2b	5	1	64/64	4	4	Conv2
Conv3	3	2	64/128	4	8	Conv2b
Conv3b	3	1	128/128	8	8	Conv3
Conv4	3	2	128/256	8	16	Conv3b
Conv4b	3	1	256/256	16	16	Conv4
Conv5	3	2	256/512	16	32	Conv4b
Conv5b	3	1	512/512	32	32	Conv5
Conv6	3	2	512/512	32	64	Conv5b
Conv6b	3	1	512/512	64	64	Conv6
Conv7	3	2	512/512	64	128	Conv6b
Conv7b	3	1	512/512	128	128	Conv7

Depth Decoder

layer	k	s	chs	in	out	input
Unconv7	3	2	512/312	128	64	Conv7b
Iconv7	3	1	1024/512	64	64	Upconv7+ Conv6b
Upconv6	3	2	512/512	64	32	Iconv7
Iconv6	3	1	1024/512	32	32	Upconv6+ Conv5b
Upconv5	3	2	512/256	32	16	Iconv6
Iconv5	3	1	512/256	16	16	Upconv5+ Conv4b
Upconv4	3	2	256/128	16	8	Iconv5
Iconv4	3	1	128/128	8	8	Upconv4+ Conv3b
disp4	3	1	128/2	8	8	Iconv4
Upconv3	3	2	128/64	8	4	Iconv4
Iconv3	3	1	130/64	4	4	Upconv3+ Conv2b+disp4
disp3	3	1	64/2	4	4	Iconv3
Upconv2	3	2	64/32	4	2	Iconv3
Iconv2	3	1	66/32	2	2	Upconv2+ Conv1b+disp3
disp2	3	1	32/2	2	2	Iconv2
Upconv1	3	2	32/16	2	1	Iconv2
Iconv1	3	1	18/16	1	1	Upconv1+disp2
disp1	3	1	16/2	1	1	Iconv1

Segmentation decoder

layer	k	s	chs	in	out	input
Unconv7	3	2	512/312	128	64	Conv7b
Iconv7	3	1	1024/512	64	64	Upconv7+Conv6b
Upconv6	3	2	512/512	64	32	Iconv7
Iconv6	3	1	1024/512	32	32	Upconv6+Conv5b
Upconv5	3	2	512/256	32	16	Iconv6
Iconv5	3	1	512/256	16	16	Upconv5+Conv4b
Upconv4	3	2	256/128	16	8	Iconv5
Iconv4	3	1	128/128	8	8	Upconv4+Conv3b
Upconv3	3	2	128/64	8	4	Iconv4
Iconv3	3	1	130/64	4	4	Upconv3+Conv2b
Upconv2	3	2	64/32	4	2	Iconv3
Iconv2	3	1	66/32	2	2	Upconv2+Conv1b
Upconv1	3	2	32/30	2	1	Iconv2
Iconv1	3	1	30/30	1	1	Upconv1

Evaluation matrix

The primary aim of our project was to verify that different applications of visual learning share the same low level features, so for evaluation we initialize the shared encoder and the depth decoder with the pre trained weights given by the author which generates state of the art results compared to other depth estimation methods. For validating the segmentation pipeline, we use the trained segmentation decoder and calculate the meanIOU over the test set of 3475 ground truth images.

One of the significant results that we achieved was on the car segmentation class where the IOU was 0.68 which is an encouraging result to explore more of such hybrid networks. Some of the other class such as motorcycle, truck, traffic signs get low score of IOU, we believe that the such poor results occur because much significant information is lost in the coarse annotation of ground truth labels which leads to poor detection of these less occurring or small classes.

Class	**IoU (Intersection over Union)**
Road	0.87
Sky	0.68
Car	0.68
Vegetation	0.62
Building	0.58
Bus	0.42
Traffic Sign	0.29
Traffic Light	0.25
wall	0.26
fence	0.22
Pole	0.23
wall	0.26
terrain	0.38
rider	0.22
truck	0.28
bus	0.42
train	0.2
motorcycle	0.14

Output

Fig: Segmentation mask

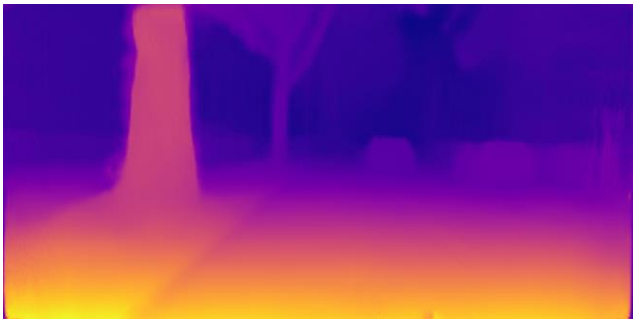


Fig: Depth map

Fig: Original image

Conclusions

In this paper the methodology of estimating depth maps and segmentation masks from a single image using a unified network has been proposed. Thus the main goal is of obtaining better hybrid CNNs that makes distinct feature extraction for a specific task and a common feature extraction for a common shared task thus modularizing the feature extraction process. Thus both the tasks get benefited from the feature extraction process without being affected by the features that are relevant only to one task thus leading to a much better performance. It was also proved that solving tasks such as semantic segmentation and

depth estimation that are correlated improves performance of individually tackling the tasks.

Future work

The future work focuses on designing a better loss function. The loss functions employed are a linear combination of loss functions of a single task. But it is very likely that these layers may have a very different meaning regarding to their physical tasks e.g. cross entropy and Euclidian loss making their unification difficult. Thus, we can find a higher-level evaluation metric which could help us in defining the loss function of the multi task system in an even better way. E.g. evaluating the prediction of 3D oriented bounding boxes on objects requires using depth and semantic segmentation results thus naturally combining the loss functions for both tasks.

References

[1] C. Godard, O. Aodha, G. Brostow: Unsupervised Monocular Depth Estimation with Left-Right Consistency

[2] R. Garg, V. Kumar BG, and I. Reid. Unsupervised CNN for
single view depth estimation: Geometry to the rescue. In ECCV, 2016.

[3] L. Ladick`y, J. Shi, and M. Pollefeys. Pulling things out of
perspective. In CVPR, 2014.

[4] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. PAMI, 2015

[5] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In NIPS, 2014

[6] G. Yang, H. Zhao, J. Shi, Z. Deng and J. Jia: SegStereo: Exploiting Semantic Information for Disparity Estimation. ECCV 2018

[7] J. Bolte, M. Kamp, A. Breuer, S. Homoceanu, P. Schlicht, F. Hüger, D. Lipinski and T. Fingscheidt: Unsupervised Domain Adaptation to Improve Image Segmentation Quality Both in the Source and Target Domain. Proc. of CVPR - Workshops 2019

[8] J. Jiao, Y. Wei∗ , Z. Jie, H. Shi, R. Lau, T. Huang: Geometry-Aware Distillation for Indoor Semantic Segmentation, CVPR 2019

[9] G. Giannone1, B. Chidlovskii: Learning Common Representation from RGB and Depth Images, CVPR 2019

[10] T. Hung, V. Jain, M. Bucher, M. Cord, P. Perez, DADA: Depth-Aware Domain Adaptation in Semantic Segmentation, CVPR 2019

[11] W. Zhuo , M. Salzmann , X. He , M. Liu: Indoor Scene Structure Analysis for Single Image Depth Estimation, CVPR 2015