Improved Monocular Depth Estimation using Semantic Information

Pratyush Kumar Sahoo [1] **Karter Krueger** [1] **Avnish Gupta** [1] **Krutarth Trivedi** [1]

Abstract

Estimating the depth of an environment from a single camera, known as monocular depth estimation, has been of growing interest as a way to replace expensive sensors on autonomous systems. We focus on improving monocular depth accuracy by exploring the relationship between depth and semantic information with multi-task learning. Our experimental results on the Cityscapes dataset demonstrate improved performance over the baseline Monodepth work.

1. Introduction

Autonomous systems such as self-driving cars and robots, require 3D depth scanning to help understand and navigate their environments. Traditionally, depth information is collected using expensive sensors, such as 3D LiDAR scanners or calibrated stereo imaging systems. In recent years, however, there has been a growing focus on estimating depth maps using a single low-cost RGB camera.

One traditional single-camera approach has relied on a sequence of images to geometrically recover 3D information across frames rather than a single image. This version, called Structure from Motion (SfM), has been applied in both 3D reconstruction and SLAM mapping successfully. The depth of sparse features is determined by SfM through feature correspondences and geometric constraints between high-quality image sequences. One downfall of SfM methods is the inability for it to determine a real depth scale.

More recent research has also explored using deep learning for the monocular depth estimation problem. This problem has been shown as an extremely challenging problem due to several issues, including low-texture environments, object occlusions, and varied lighting with shadows. One issue commonly seen due to lack of texture is objects blending into the background without distinct edges. Blurred edges is a clear problem that could lead an autonomous system to collide with an obstacle. We explore the intuition that including object semantics/class segmentation during training may improve depth estimation. We expect this to teach the network how pixel-level object class information affects features used for estimating depth. We show how this leads to more distinctive object edges in the disparity map compared to existing methods.

For this multi-task network, we require both the left and right images, and the pixel-wise semantically-labeled ground truth images. For this reason, we chose to use the Cityscapes (Cordts et al., 2016) dataset which has semantic labels for both left and right images. Many other popular datasets, such as KITTI (Geiger et al., 2012), include left and right images, and ground-truth depth data, but not semantics.

1.1. Research contributions

The core contribution of this paper comes from combining monocular depth estimation and semantic segmentation into one multi-task network to improve depth estimation accuracy. We extended the existing Monodepth network by adding a second decoder that is dedicated to the semantic segmentation task. The dual-decoder, multi-task network is trained with a hybrid loss function that combines the original depth-disparity losses from the depth decoder with the softmax cross-entropy loss of the semantic decoder. We also conduct experiments with an additional Dice loss added to the semantic decoder loss, with additional improvements shown.

2. Related Work

Prior work in depth estimation includes research in methods using stereo-pairs, leveraging multi-view scene geometry, and temporal image sequences. Monocular depth estimation has been treated as both a supervised and self-supervised learning problem. We briefly review the classic stereo approach as well as the more recent monocular approaches.

[1]Worcester Polytechnic Institute. Correspondence to: Pratyush Kumar Sahoo <psahoo@wpi.edu>, Karter Krueger <kkrueger2@wpi.edu>, Avnish Gupta <agupta8@wpi.edu>, Krutarth Trivedi <ktrivedi@wpi.edu>.

CS/DS 541 Deep Learning Final Project, Worcester Polytechnic Institute.

2.1. Classical Approach

The classical approach to estimating the depth of a scene using visual data is to use two cameras configured in a synchronized stereo setup with a known baseline distance between lenses. The two frames, with known spacing, can be used to geometrically determine the distance of image key-points by means of the triangulation method where the light rays intersect the view plane in the pin-hole camera model.

2.2. Supervised Single Image Depth Estimation

Many supervised learning methods have achieved significant improvements in finding a relationship between color images and their corresponding depth maps using convolutional networks. These methods generally require ground truth labels for training.

Eigen et al. (Eigen et al., 2014) showed the possibility of producing a dense pixel depth map by using a two-scale deep network trained on images to predict corresponding depth values utilizing a coarse-scale network to predict the global scene depth and consequently a local level fine-scale network that edits the coarse prediction to align with local details such as object edges. Ladicky et al. (Ladicky et al., 2014) incorporated semantics to perform joint classification on pixel-wise depth and semantic labels, but suffered from model bias towards objects present at specific depth values in the dataset.

2.3. Self-Supervised Depth Estimation

In the absence of ground truth depth data, one alternative is to train depth estimation models using image reconstruction or Structure from Motion(SfM) as the supervisory signal to generate a depth map. In the case of stereo image pairs, the depth can be estimated by finding the disparity between them. In the case of a monocular camera, the depth can be estimated using image reconstruction by using the network to estimate the camera motion. The training can then be performed in a self-supervised manner using the estimated movement between frames.

Monodepth1 (Godard et al., 2016) addressed the problem of monocular depth estimation by using a deep network to generate the corresponding right viewpoint from an input left image to then estimate disparity similar to a stereo camera. By posing monocular depth estimation as an image reconstruction problem, they find the disparity field without requiring ground truth depth. However, only minimizing a photometric loss resulted in good quality image reconstructions but poor quality depth. Hence, they added a left-right consistency loss to improve the quality of the synthesized depth images.

The extended version of Monodepth1, named Monodepth2 (Godard et al., 2018), proposed a network to predict the appearance of a target image from the viewpoint of another image by performing image synthesis. Similar to Monodepth1, they expressed the relative pose for each source viewpoint with respect to the target viewpoint pose and predicted a dense depth map that minimizes the photometric reprojection error between the frames.

2.4. Cityscapes

Cityscapes (Cordts et al., 2016) is a large-scale database that focuses on semantic understanding of urban street scenes. It provides semantic, instance-wise, and dense pixel annotations for 30 classes grouped into 8 different categories such as flat surfaces, humans, vehicles, constructions, objects, nature, sky, and void. The dataset consists of around 5000 fine annotated images and 20,000 coarse annotated ones captured in 50 cities during several months, daytimes, and good weather conditions.

3. Proposed Method

We focus on reusing the original VGG16 depth encoder and decoder from Monodepth, but with an added semantic decoder that branches from the same encoder. The motivation behind this is related to the work done with a MultiTask architecture with two decoders in (Ramirez et al., 2018) and the thought that adding a semantic decoder will force the encoder network to learn a better representation of object edges in the scene and lead to improved depth output from the original decoder.

We attach a second decoder inspired by the UNet semantic segmentation (Ronneberger et al., 2015) architecture. This decoder follows a similar architecture to the depth decoder that progressively scales the feature space back up through layers of up-convolutions and up-sampling with Skip connections from the encoder. The purpose of adding skip connections is to preserve some features from the larger image that were encoded at each layer on the way down to the latent space in the middle. The number of convolutional channels in the layers of the decoder goes: 512, 512, 256, 128, 64, 64, 34. The final layer being 34 channels, which is due to the 34 semantic classes of the Cityscapes dataset (Cordts et al., 2016). After the final layer, a softmax activation is applied along the channel axis for every pixel representing probability scores of a pixel belonging to a semantic class.

The main parameters that we adjusted were the number of convolutional channels of the added semantic decoder. The channels originally would go down to filters of 32, 16, and then back up to the final layer of 34, but it seemed as if this was limiting the ability to learn the final classes, so

we changed the three final layers to go 64, 64, 64, and then to the 34 classes. This was effective as it showed improvements to learn semantics. We also adjusted the loss function (explained in more detail below) to include different weightings of the depth vs semantic loss subtotals. After several trials, we found it was best to weight the depth losses and semantic loss both at a 1.0 multiplier, as the depth loss would inherently be considered more important thanks to the 4 sub losses being summed in it's total first, which cause it to fall in the 0-4 range. The semantic softmax loss typically ranged in the 0-1.4 range and was considered less important over depth (which was our desired metric to improve). We experimented with adding a Dice loss into the semantic subtotal and found it to improve underrepresented classes, so this was averaged in with softmax loss.

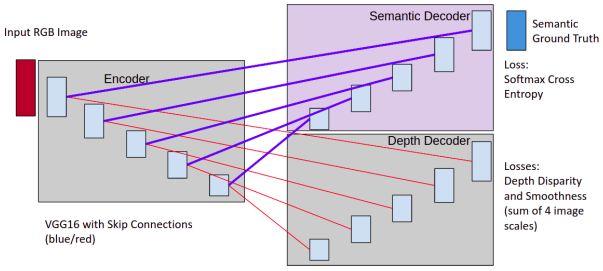


Figure 1. Network architecture with an added semantic decoder. The loss combines depth losses with semantic loss.

3.1. Training Loss

We formulate a multi-task loss that is the weighted sum of the depth-disparity loss calculated across four scales and the semantic-segmentation loss calculated by softmax-cross entropy across the channel dimension of each pixel. The depth-disparity loss for each of the four scales is computed as a combination of three terms,

$$C_d = \sum_{i=1}^{S} (\alpha_{ap}(C_{ap}) + \alpha_{ds}(C_{ds}) + \alpha_{lr}(C_{lr})) \quad (1)$$

Each of the losses above has a left and a corresponding right term, C_{ap} is the term which measures similarity between the reconstructed image and the original using the disparity and encourages the reconstructed image to match the original. C_{ap} is the combination of the L1 and Single Scale SSIM (Wang et al., 2004) loss.

C_{ds} denotes the disparity smoothness loss that encourages smoother depth output by penalizing disparity gradients including an edge-aware term which caters to the relation of larger image gradients and depth discontinuities.

C_{lr} denotes the left-right consistency loss that enhances the accuracy of disparity prediction by using the network to output both the left and the right disparity maps and forces them to be consistent by minimizing an L1 loss between both the disparity maps.

C_{sm} denotes the softmax cross-entropy loss for semantic segmentation. We also added Dice Loss C_{sm} which improves segmentation accuracy for classes which are less in occurrence. The final training loss is calculated in Equation 2. The w_d and w_{sm} are the corresponding weight coefficients for the losses.

$$C_{total} = w_d(C_d) + w_{sm}(C_{sm}) \quad (2)$$

4. Experiments

Here, we validate that our multi-task approach improves depth accuracy by learning from the estimated semantic information during training. We evaluate our models on the Cityscapes dataset (Cordts et al., 2016) to allow comparison with the existing monocular methods.

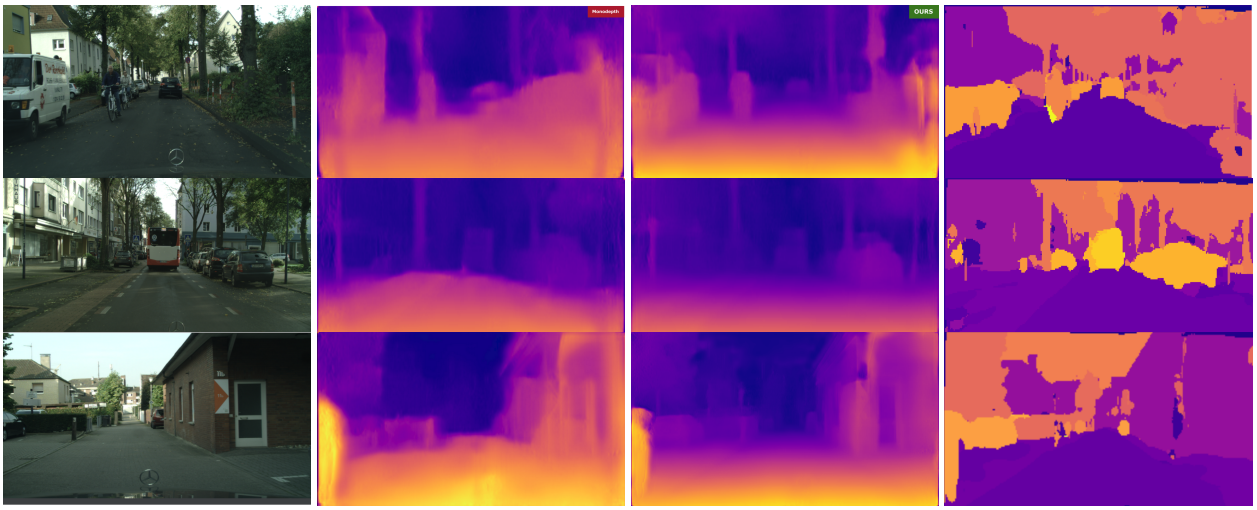
We experimented with several hyperparameters during training time. We started with a batch size of 1 due to GPU constraints, but found this to have very noisy loss values and wasn't converging. We then increased this to batch size of 2, and then later 4, by lowering the input resolution to 128x256 to still fit on the GPU memory. The batch size 4 performed the best and led to loss values as low as 0.15 for depth and 0.19 for semantics as shown in Figure 3. After further testing, we decided it was important to go back to a larger 256x512 size of image, which was possibly by adjusting the dataloader to pre-load only 128 images instead of the standard 2048 (which reduced training speed).

We trained our model on Jetson Xavier AGX with 32GB RAM shared between GPU and CPU. We used the learning rate of 1e-4 (same as original Monodepth) and a batch size of 4 and the Adam optimizer. We used Cityscapes refined dataset with 5000 labeled images of which 3750 are used in the training set. We trained for approximately 180,000 batches, which is 192 epochs and it took around twenty four hours of training. The network was trained for two different input image resolution 128 x 256 and 256 x 512, with results shown for both and the latter performing the best.

5. Results

The images in Figure 2 shows the result of our implementation against the original Monodepth1 at different image resolutions. It is evident from the results that semantic information has helped improve the depth quality by giving information about object boundaries. Specifically for pixels of the road class have a smoother disparity as constrained by semantic information in our implementation.

Our approach was trained on the cityscapes dataset with standard benchmarks then performed on the KITTI Eigen



(i) Original Image (j) Monodepth1 - Depth (k) Ours - Depth (l) Ours - Semantic

Figure 2. Our method achieves superior qualitative results on Cityscapes. First two rows show result for input image size of 128 x 256 and the last row shows result for input image size 256 x 512.

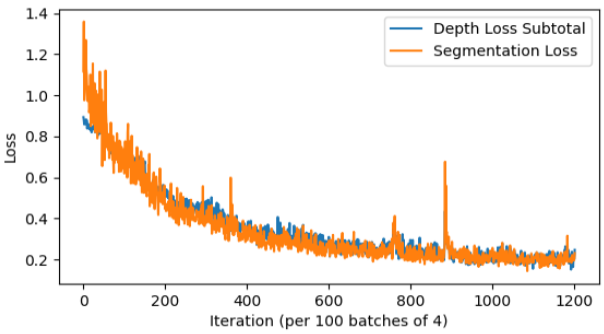


Figure 3. Loss curve for total depth loss and segmentation loss

test set. Our results show comparable quantitative results, although slightly worse than the baseline Monodepth1, as shown in the appendix table.

6. Discussion

Monodepth1 is a pioneer work that introduced unsupervised training for depth estimation using a single image, however, we found that their network fails to estimate depth boundaries accurately for low-texture objects. We confirmed our hypothesis that adding semantics during training would improve the distinction of objects, as seen in Fig2.

The features representing semantic information in the common encoder backbone helps the depth decoder to understand the uniform depth gradient for pixels belonging to the same regular object which encourages our network to output smoother disparity maps as demonstrated by our smooth ground plane in Figure 2 as well.

We also observed that our network fails in case of distant objects, which can be attributed to low training resolution. The depth decoder also struggled to contain depth values for some objects which could be due to incorrect pixel segmentation that leads it to perceive a false structure.

7. Conclusions and Future Work

We have presented a multi-task convolutional neural network architecture for single image depth prediction employing semantic information. The joint optimization helps the convolutional neural network to learn better disparity representations using a single image. We conclude that forcing the network to optimize over semantic pixel-wise constraints of the scene alleviates the ambiguities in monocular depth estimation consequently encouraging better depth predictions. The proposed network predicts smoother disparity maps by reasoning crisp edges for objects in the disparity maps preserving structure. Through extensive experimental evaluations on the CityScapes dataset, the proposed network achieves outstanding performance over the baseline, Monodepth1.

The results from our experiments encourage future research into further multi-task learning semantic features. We observe that the semantic decoder affects the depth decoder performance successfully. There are some cases where the semanatics are inaccurate, so we believe future work could focus on adding losses for scaled semantic ground truth at each feature map, which would make it more scale invariant and would improve semantic predictions. We are also experimenting with adding a pre-trained semantic encoder block hoping to use pre-trained semantic feature maps.

8. Appendix

Evaluation results				
Network	bs$_r el$	sq$_r el$	rms	log$_r ms$
Mono1	1.1962	38.9447	20.577	0.764
Ours	1.4175	50.4493	24.561	0.86

We compare the monodepth pre-trained model trained on 30,000 images from the cityscapes dataset with our semantic decoder approach which we train on an 3750 images with ground-truth semantic labels on the KITTI-Eigen split having 690 images. We see that we achieve comparabale results on all the accuracy metrics including abs-rel, sqr-rel, rmse and log-rmse. We also obtain much smoother disparity maps which is a better representation of the depth disparity gradients, our work motivates and indicates that training on more data having semantic information we would be potentially outperforming the Monodepth-1 benchmark.

References

Cordts, Marius, Omran, Mohamed, Ramos, Sebastian, Rehfeld, Timo, Enzweiler, Markus, Benenson, Rodrigo, Franke, Uwe, Roth, Stefan, and Schiele, Bernt. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Eigen, David, Puhrsch, Christian, and Fergus, Rob. Depth map prediction from a single image using a multi-scale deep network, 2014. URL https://arxiv.org/abs/1406.2283.

Geiger, Andreas, Lenz, Philip, and Urtasun, Raquel. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Godard, Clément, Aodha, Oisin Mac, and Brostow, Gabriel J. Unsupervised monocular depth estimation with left-right consistency. *CoRR*, abs/1609.03677, 2016. URL http://arxiv.org/abs/1609.03677.

Godard, Clément, Aodha, Oisin Mac, and Brostow, Gabriel J. Digging into self-supervised monocular depth estimation. *CoRR*, abs/1806.01260, 2018. URL http://arxiv.org/abs/1806.01260.

Ladicky, L., Shi, Jianbo, and Pollefeys, M. Pulling things out of perspective. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, June 2014. doi: 10.1109/CVPR.2014.19.

Ramirez, Pierluigi Zama, Poggi, Matteo, Tosi, Fabio, Mattoccia, Stefano, and di Stefano, Luigi. Geometry meets semantics for semi-supervised monocular depth estimation. *CoRR*, abs/1810.04093, 2018. URL http://arxiv.org/abs/1810.04093.

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Wang, Zhou, Bovik, A.C., Sheikh, H.R., and Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.