AuxNet: Auxiliary tasks enhanced
Semantic Segmentation for Automated Driving

Sumanth Chennupati[1,3], Ganesh Sistu[2], Senthil Yogamani[2] and Samir Rawashdeh[3]

[1]*Valeo Troy, United States*

[2]*Valeo Vision Systems, Ireland*

[3]*University of Michigan-Dearborn*

{*sumanth.chennupati, ganesh.sistu,senthil.yogamani*}@*valeo.com, srawa@umich.edu*

Keywords: Semantic Segmentation, Multitask Learning, Auxiliary Tasks, Automated Driving.

Abstract: Decision making in automated driving is highly specific to the environment and thus semantic segmentation plays a key role in recognizing the objects in the environment around the car. Pixel level classification once considered a challenging task which is now becoming mature to be productized in a car. However, semantic annotation is time consuming and quite expensive. Synthetic datasets with domain adaptation techniques have been used to alleviate the lack of large annotated datasets. In this work, we explore an alternate approach of leveraging the annotations of other tasks to improve semantic segmentation. Recently, multi-task learning became a popular paradigm in automated driving which demonstrates joint learning of multiple tasks improves overall performance of each tasks. Motivated by this, we use auxiliary tasks like depth estimation to improve the performance of semantic segmentation task. We propose adaptive task loss weighting techniques to address scale issues in multi-task loss functions which become more crucial in auxiliary tasks. We experimented on automotive datasets including SYNTHIA and KITTI and obtained 3% and 5% improvement in accuracy respectively.

scale

1 Introduction

Semantic image segmentation has witnessed tremendous progress recently with deep learning. It provides dense pixel-wise labeling of the image which leads to scene understanding. Automated driving is one of the main application areas where it is commonly used. The level of maturity in this domain has rapidly grown recently and the computational power of embedded systems have increased as well to enable commercial deployment. Currently, the main challenge is the cost of constructing large datasets as pixel-wise annotation is very labor intensive. It is also difficult to perform corner case mining as it is a unified model to detect all the objects in the scene. Thus there is a lot of research to reduce the sample complexity of segmentation networks by incorporating domain knowledge and other cues wherever possible. One way to overcome this is via using synthetic datasets and domain adaptation techniques (Sankaranarayanan et al., 2018), another way is to use multiple clues or annotations to learn efficient representations for the task with limited or expensive annotations (Liebel and Körner, 2018).

Figure 1: Semantic Segmentation of an automotive scene

In this work, we explore the usage of auxiliary task learning to improve the accuracy of semantic segmentation. We demonstrate the improvements in semantic segmentation by inducing depth cues via auxiliary learning of depth estimation. The closest related work is (Liebel and Körner, 2018) where auxiliary task was used to improve semantic segmentation task using GTA game engine. Our work demonstrates it for real and synthetic datasets using novel loss functions. The contributions of this work include:

1. Construction of auxiliary task learning architecture for semantic segmentation.

2. Novel loss function weighting strategy for one main task and one auxiliary task.

3. Experimentation on two automotive datasets namely KITTI and SYNTHIA.

The rest of the paper is organized as follows: Section 2 reviews the background in segmentation in automated driving and learning using auxiliary tasks. Section 3 details the construction of auxiliary task architecture and proposed loss function weighting strategies. Section 4 discusses the experimental results in KITTI and SYNTHIA. Finally, section 5 provides concluding remarks.

2 Background

2.1 Semantic Segmentation

A detailed survey of semantic segmentation for automated driving is presented in (Siam et al., 2017). We briefly summarize the relevant parts focused on CNN based methods. FCN (Long et al., 2015) is the first CNN based end to end trained pixel level segmentation network. Segnet (Badrinarayanan et al., 2017) introduced encoder decoder style semantic segmentation. U-net (Çiçek et al., 2016) is also an encoder decoder network with dense skip connections between the them. While these papers focus on architectures, Deeplab (Chen et al., 2018a) and EffNet (Freeman et al., 2018) focused on efficient convolutional layers by using dilated and separable convolutions.

Annotation for semantic segmentation is a tedious and expensive process. An average experienced annotator takes anywhere around 10 to 20 minutes for one image and it takes 3 iterations for correct annotations, this process limit the availability of large scale accurately annotated datasets. Popular semantic segmentation automotive datasets like CamVid (Brostow et al., 2008), Cityscapes (Cordts et al., 2016), KITTI (Geiger et al., 2013) are relatively smaller when compared to classification datasets like ImageNet (Deng et al., 2009). Synthetic datasets like Synthia (Ros et al., 2016), Virtual KITTI (Gaidon et al., 2016), Synscapes (Wrenninge and Unger, 2018) offer larger annotated synthetic data for semantic segmentation. Efforts like Berkley Deep Drive (Xu et al., 2017), Mapillary Vistas (Neuhold et al., 2017) and Toronto City (Wang et al., 2017) have provided larger datasets to facilitate training a deep learning model for segmentation.

2.2 Multi-Task Learning

Multi-task learning (Kokkinos, 2017), (Chen et al., 2018b), (Neven et al., 2017) has been gaining significant popularity over the past few years as it has proven to be very efficient for embedded deployment. Multiple tasks like object detection, semantic segmentation, depth estimation etc can be solved simultaneously using a single model. A typical multi-task learning framework consists of a shared encoder coupled with multiple task dependent decoders. An encoder extracts feature vectors from an input image after series of convolution and poling operations. These feature vectors are then processed by individual decoders to solve different problems. (Teichmann et al., 2018) is an example where three task specific decoders were used for scene classification, object detection and road segmentation of an automotive scene. The main advantages of multi-task learning are improved computational efficiency, regularization and scalability. (Ruder, 2017) discusses other benefits and applications of multi-task learning in various domains.

2.3 Auxiliary Task Learning

Learning a side or auxiliary task jointly during training phase to enhance main task's performance is usually referred to auxiliary learning. This is similar to multi-task learning except the auxiliary task is non-operational during inference. This auxiliary task is usually selected to have much larger annotated data so that it acts a regularizer for main task. In (Liebel and Körner, 2018) semantic segmentation is performed using auxiliary tasks like time, weather, etc. In (Toshniwal et al., 2017), end2end speech recognition training uses auxiliary task phoneme recognition for initial stages. (Parthasarathy and Busso, 2018) uses unsupervised aux tasks for audio based emotion recognition. It is often believed that auxiliary tasks can be used to focus attention on a specific parts of the input. Predictions of road characteristics like markings as an auxiliary task in (Caruana, 1997) to improve main task for steering prediction is one instance of such behaviour.

Figure 2 illustrates auxiliary tasks in a popular automated driving dataset KITTI. It contains various dense output tasks like Dense optical flow, depth estimation and visual SLAM. It also contains meta-data like steering angle, location and external condition. These meta-data comes for free without any annotation task. Depth could be obtained for free by making use of Velodyne depth map, (Kumar et al., 2018) demonstrate training using sparse Velodyne supervision.

```
data['location']  = Urban Street
data['time']      = Day
data['condition'] = Sunny
imu['speed']      = 20mph
imu['steering']   = +10°
Imu['direction']  = S'E'
```
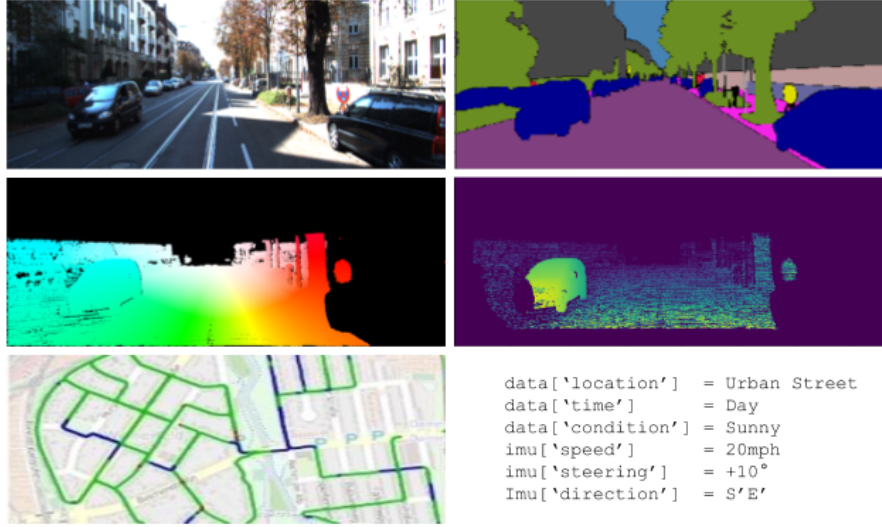
Figure 2: Illustration of several Auxiliary Visual Perception tasks in an Automated driving dataset KITTI. First Row shows RGB and Semantic Segmentation, Second Row shows Dense Optical Flow and Depth, Third row shows Visual SLAM and meta-data for steering angle, location and condition

2.4 Multi-Task Loss

Modelling a multi-task loss function is a critical step in multi-task training. An ideal loss function should enable learning of multiple tasks with equal importance irrespective of loss magnitude, task complexity etc. Manual tuning of task weights in a loss function is a tedious process and it is prone to errors. Most of the work in multi-task learning uses a linear combination of multiple task losses which is not effective. (Kendall et al., 2018) propose an approach to learn the optimal weights adaptively based on uncertainty of prediction. The log likelihood of the proposed joint probabilistic model shows that the task weights are inversely proportional to the uncertainty. Minimization of total loss w.r.t task uncertainties and losses converges to an optimal loss weights distribution. This enables independent tasks to learn at a similar rate allowing each to influence on training. However, these task weights are adjusted at the beginning of the training and are not adapted during the learning. GradNorm (Chen et al., 2018c) proposes an adaptive task weighing approach by normalizing gradients from each task. They also consider the rate of change of loss to adjust task weights. (Liu et al., 2018) adds a moving average of task weights obtained by method similar to GradNorm. (Guo et al., 2018) on other hand proposes dynamic weight adjustments based on task difficulty. As the difficulty of learning changes over training time, the task weights are updated allowing the model to prioritize difficult tasks. Modelling multi-task loss as a multi-objective function was proposed in (Zhang and Yeung, 2010), (Sener and Koltun, 2018) and (Désidéri, 2009). A reinforcement learning approach was used in (Liu, 2018) to minimize the total loss while changing the loss weights.

3 Methods

Semantic segmentation and depth estimation have common feature representations. Joint learning of these tasks have shown significant performance gains in (Liu et al., 2018), (Eigen and Fergus, 2015), (Mousavian et al., 2016), (Jafari et al., 2017) and (Gurram et al., 2018). Learning underlying representations between these tasks help the multi-task network alleviate the confusion in predicting semantic boundaries or depth estimation. Inspired by these papers, we propose a multi-task network with semantic segmentation as main task and depth estimation as an auxiliary task. As accuracy of the auxiliary task is not important, weighting its loss function appropriately is important. We also discuss in detail the proposed auxiliary learning network and how we overcame the multi-task loss function challenges discussed in section 2.4.

3.1 Architecture Design

The proposed network takes input RGB image and outputs semantic and depth maps together. Figure 3 shows two task specific decoders coupled to a shared encoder to perform semantic segmentation and depth
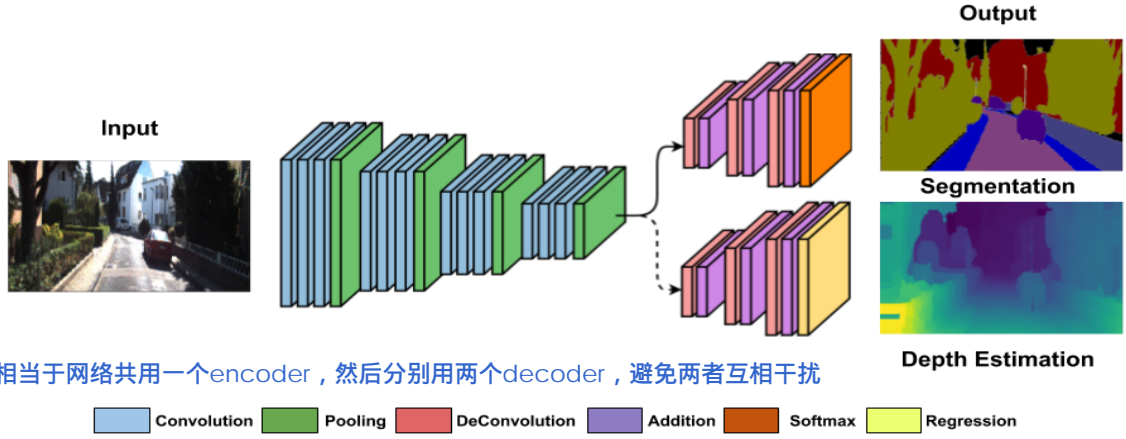
Figure 3: AuxNet: Auxiliary Learning network with Segmentation as main task and Depth Estimation as auxiliary task.

estimation. The shared encoder is built using ResNet-50 (He et al., 2016) by removing the fully connected layers from the end. The encoded feature vectors are now passed to two parallel stages for independent task decoding. Semantic segmentation decoder is constructed similar to FCN8 (Long et al., 2015) architecture with transposed convolutions, up sampling and skip connections. The final output is made up of a softmax layer to output probabilistic scores for each semantic class. Depth estimation decoder is also constructed similar to segmentation decoder except the final output is replaced with a regression layer to estimate scalar depth.

3.2 Loss Function

In general, a multi-task loss function is expressed as weighted combination of multiple task losses where $\mathcal{L}_i$ is loss and λ_i is associated weight for task i.

$$\mathcal{L}_{Total} = \sum_{i=1}^{t} \lambda_i \mathcal{L}_i \qquad (1)$$

For the proposed 2-task architecture we express loss as:

$$\mathcal{L}_{Total} = \lambda_{Seg} \mathcal{L}_{Seg} + \lambda_{Depth} \mathcal{L}_{Depth} \qquad (2)$$

$\mathcal{L}_{Seg}$ is semantic segmentation loss expressed as an average of pixel wise cross-entropy for each predicted label and ground truth label. $\mathcal{L}_{Depth}$ is depth estimation loss expressed as mean absolute error between estimated depth and true depth for all pixels. To overcome the significant scale difference between semantic segmentation and depth estimation losses, we perform task weight balancing as proposed in Algorithm 1. Expressing multi-task loss function as product of task losses, forces each task to optimize so that the total loss reaches a minimal value. This ensures no

task is left in a stale mode while other tasks are making progress. By making an update after every batch in an epoch, we dynamically change the loss weights. We also add a moving average to the loss weights to smooth the rapid changes in loss values at the end of every batch.

> **for** $epoch \leftarrow 1$ **to** n **do**
>   **for** $batch \leftarrow 1$ **to** s **do**
> $\lambda_{Seg} \quad = \mathcal{L}_{Depth}$
> $\lambda_{Depth} = \mathcal{L}_{Seg}$
> $\mathcal{L}_{Total} = \mathcal{L}_{Depth} \mathcal{L}_{Seg} + \mathcal{L}_{Seg} \mathcal{L}_{Depth}$
> $\mathcal{L}_{Total} = 2 \times \mathcal{L}_{Seg} \mathcal{L}_{Depth}$
>   **end**
> **end**

Algorithm 1: Proposed Weight Balancing for 2-task semantic segmentation and depth estimation

In Algorithm 2, we propose focused task weight balancing to prioritize the main task's loss in auxiliary learning networks. We introduce an additional term to increase the weight of main task. This term could be a fixed value to scale up main task weight or the magnitude of task loss.

> **for** $epoch \leftarrow 1$ **to** n **do**
>   **for** $batch \leftarrow 1$ **to** s **do**
> $\lambda_{Seg} \quad = \mathcal{L}_{Seg} \times \mathcal{L}_{Depth}$
> $\lambda_{Depth} = \mathcal{L}_{Seg}$
> $\mathcal{L}_{Total} = \mathcal{L}_{Seg}^2 \mathcal{L}_{Depth} + \mathcal{L}_{Seg} \mathcal{L}_{Depth}$
> $\mathcal{L}_{Total} = (\mathcal{L}_{Seg} + 1) \times \mathcal{L}_{Seg} \mathcal{L}_{Depth}$
>   **end**
> **end**

Algorithm 2: Proposed Focused Task Weight Balancing for Auxiliary Learning.
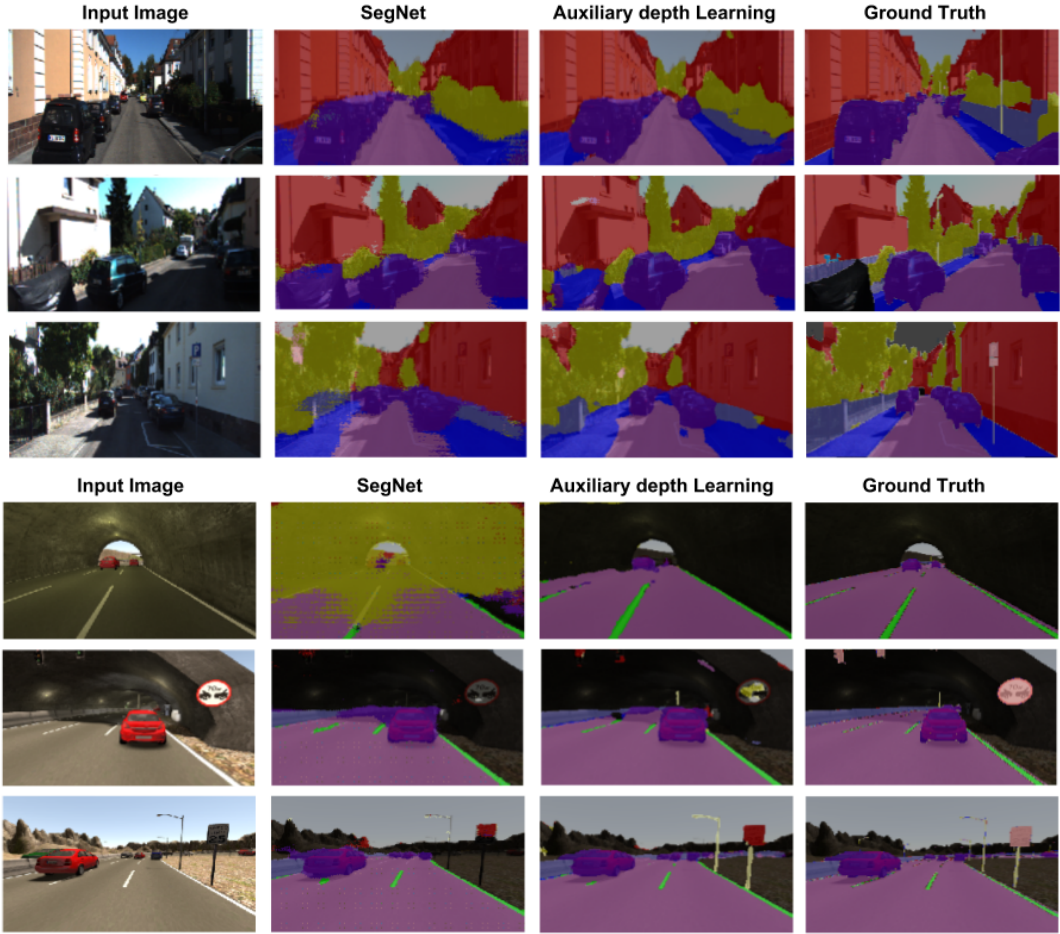
Figure 4: Results on KITTI and SYNTHIA datasets

4 Results and Discussion

In this section, we present details about the experimental setup used and discuss the observations on the results obtained.

4.1 Experimental Setup

We implemented the auxiliary learning network as discussed in section 3.1 to perform semantic segmentation and depth estimation. We chose ResNet-50 as the shared encoder which is pre-trained on ImageNet. We used segmentation and depth estimation decoders with random weight initialization. We performed all our experiments on KITTI (Geiger et al., 2013) semantic segmentation and SYNTHIA (Ros et al., 2016) datasets. These datasets contain RGB image data, ground truth semantic labels and depth data represented as disparity values in 16-bit png format. We re-sized all the input images to a size 224x384.

The loss function is expressed as detailed in section 3.2. Categorical cross-entropy was used to compute semantic segmentation loss and mean absolute error is used to compute depth estimation loss. We implemented four different auxiliary learning networks by changing the expression of loss function. $AuxNet_{400}$ and $AuxNet_{1000}$ weighs segmentation loss 400 and 1000 times compared to depth estimation loss. $AuxNet_{TWB}$ and $AuxNet_{FTWB}$ are built based on Algorithms 1 and 2 respectively. These networks are trained with ADAM (Kingma and Ba, 2014) optimizer for 200 epochs. The best model for each network was saved by monitoring the validation loss of semantic segmentation task. Mean IoU and categorical IoU were used for comparing the performance.

4.2 Results and Discussion

In Table 1, we compare the proposed auxiliary learning networks (AuxNet) against a simple semantic segmentation network (SegNet) constructed using an en-

Table 1: Comparison Study : SegNet vs different auxiliary networks.

KITTI									
Model	**Sky**	**Building**	**Road**	**Sidewalk**	**Fence**	**Vegetation**	**Pole**	**Car**	**Lane**	**IoU**
SegNet	46.79	87.32	89.05	60.69	**22.96**	85.99	-	74.04	-	74.52
$AuxNet_{400}$	49.11	88.55	93.17	69.65	22.93	87.12	-	74.63	-	78.32
$AuxNet_{1000}$	49.17	89.81	90.77	64.16	14.77	86.52	-	71.40	-	76.58
$AuxNet_{TWB}$	**49.73**	**91.10**	92.30	70.55	18.64	86.01	-	77.32	-	78.64
$AuxNet_{FTWB}$	48.43	89.50	**92.71**	**71.58**	15.37	**88.31**	-	**79.55**	-	**79.24**

SYNTHIA									
Model	**Sky**	**Building**	**Road**	**Sidewalk**	**Fence**	**Vegetation**	**Pole**	**Car**	**Lane**	**IoU**
SegNet	95.41	58.18	93.46	09.82	76.04	80.95	08.79	85.73	90.28	89.70
$AuxNet_{400}$	95.12	**69.82**	92.95	21.38	77.61	84.23	51.31	**90.42**	91.20	91.44
$AuxNet_{1000}$	95.41	59.57	**96.83**	28.65	**81.23**	82.48	**56.43**	88.93	94.19	**92.60**
$AuxNet_{TWB}$	94.88	66.41	94.81	**31.24**	77.01	**86.04**	21.83	90.16	**94.47**	91.67
$AuxNet_{FTWB}$	**95.82**	56.19	96.68	21.09	81.19	83.26	55.86	89.01	92.11	92.05

coder decoder combination. The main difference between these two networks is the additional depth estimation decoder. It is observed that auxiliary networks perform better than the baseline semantic segmentation. It is evident that incorporating depth information improves the performance of segmentation task. It is also observed that depth dependent categories like sky, sidewalk, pole and car have shown better improvements than other categories due to availability of depth cues.

Table 2: Comparison between SegNet, FuseNet and AuxNet in terms of performance and parameters.

KITTI		
Model	**IoU**	**Params**
SegNet	74.52	23,672,264
FuseNet	**80.99**	47,259,976
AuxNet	79.24	23,676,142

SYNTHIA		
Model	**IoU**	**Params**
SegNet	89.70	23,683,054
FuseNet	92.52	47,270,766
AuxNet	**92.60**	23,686,932

We compare the performances of SegNet, AuxNet with FuseNet in Table 2. FuseNet is another semantic segmentation network (FuseNet) that takes RGB images and depth map as input. It is constructed in a similar manner to the work in (Hazirbas et al., 2016). We compare the mean IoU of each network and the number of parameters needed to construct the network. AuxNet required negligible increase in parameters while FuseNet almost needed twice the number of parameters compared to SegNet. It is observed AuxNet can be chosen as a suitable low cost replacement to FuseNet as the needed depth information is learned by shared encoder.

5 Conclusion

Semantic segmentation is a critical task to enable fully automated driving. It is also a complex task and requires large amounts of annotated data which is expensive. Large annotated datasets is currently the bottleneck for achieving high accuracy for deployment. In this work, we look into an alternate mechanism of using auxiliary tasks to alleviate the lack of large datasets. We discuss how there are many auxiliary tasks in automated driving which can be used to improve accuracy. We implement a prototype and use depth estimation as an auxiliary task and show 5% improvement on KITTI and 3% improvement on SYNTHIA datasets. We also experimented with various weight balancing strategies which is a crucial problem to solve for enabling more auxiliary tasks. In future work, we plan to augment more auxiliary tasks.

REFERENCES

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:2481–2495.

Brostow, G. J., Fauqueur, J., and Cipolla, R. (2008). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, xx(x):xx–xx.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018a). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848.

Chen, L., Yang, Z., Ma, J., and Luo, Z. (2018b). Driving scene perception network: Real-time joint detection, depth estimation and semantic segmentation. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Chen, Z., Badrinarayanan, V., Lee, C.-Y., and Rabinovich, A. (2018c). Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Désidéri, J.-A. (2009). Multiple-gradient descent algorithm (mgda).

Eigen, D. and Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *2015 IEEE International Conference on Computer Vision (ICCV)*.

Freeman, I., Roese-Koerner, L., and Kummert, A. (2018). Effnet: An efficient structure for convolutional neural networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 6–10.

Gaidon, A., Wang, Q., Cabon, Y., and Vig, E. (2016). Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*.

Guo, M., Haque, A., Huang, D.-A., Yeung, S., and Fei-Fei, L. (2018). Dynamic task prioritization for multitask learning. In *European Conference on Computer Vision*, pages 282–299. Springer.

Gurram, A., Urfalioglu, O., Halfaoui, I., Bouzaraa, F., and Lopez, A. M. (2018). Monocular depth estimation by learning from heterogeneous datasets. *2018 IEEE Intelligent Vehicles Symposium (IV)*.

Hazirbas, C., Ma, L., Domokos, C., and Cremers, D. (2016). Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision*, pages 213–228. Springer.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Jafari, O. H., Groth, O., Kirillov, A., Yang, M. Y., and Rother, C. (2017). Analyzing modular cnn architectures for joint depth prediction and semantic segmentation. *2017 IEEE International Conference on Robotics and Automation (ICRA)*.

Kendall, A., Gal, Y., and Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

Kokkinos, I. (2017). Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5454–5463.

Kumar, V. R., Milz, S., Witt, C., Simon, M., Amende, K., Petzold, J., Yogamani, S., and Pech, T. (2018). Monocular fisheye camera depth estimation using sparse lidar supervision. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2853–2858. IEEE.

Liebel, L. and Körner, M. (2018). Auxiliary tasks in multi-task learning. *arXiv preprint arXiv:1805.06334*.

Liu, S. (2018). *EXPLORATION ON DEEP DRUG DISCOVERY: REPRESENTATION AND LEARNING*. PhD thesis, UNIVERSITY OF WISCONSIN-MADISON.

Liu, S., Johns, E., and Davison, A. J. (2018). End-to-end multi-task learning with attention.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.

Mousavian, A., Pirsiavash, H., and Kosecka, J. (2016). Joint semantic segmentation and depth estimation with deep convolutional networks. *2016 Fourth International Conference on 3D Vision (3DV)*.

Neuhold, G., Ollmann, T., Bulo, S. R., and Kontschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, pages 5000–5009.

Neven, D., Brabandere, B. D., Georgoulis, S., Proesmans, M., and Gool, L. V. (2017). Fast scene understanding for autonomous driving.

Parthasarathy, S. and Busso, C. (2018). Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes. In *Interspeech*.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3234–3243.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks.

Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S. N., and Chellappa, R. (2018). Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*.

Sener, O. and Koltun, V. (2018). Multi-task learning as multi-objective optimization.

Siam, M., Elkerdawy, S., Jagersand, M., and Yogamani, S. (2017). Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8.

Teichmann, M., Weber, M., Zollner, M., Cipolla, R., and Urtasun, R. (2018). Multinet: Real-time joint semantic reasoning for autonomous driving. *2018 IEEE Intelligent Vehicles Symposium (IV)*.

Toshniwal, S., Tang, H., Lu, L., and Livescu, K. (2017). Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition. In *INTERSPEECH*.

Wang, S., Bai, M., Máttyus, G., Chu, H., Luo, W., Yang, B., Liang, J., Cheverie, J., Fidler, S., and Urtasun, R. (2017). Torontocity: Seeing the world with a million eyes. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3028–3036.

Wrenninge, M. and Unger, J. (2018). Synscapes: A photorealistic synthetic dataset for street scene parsing. *CoRR*, abs/1810.08705.

Xu, H., Gao, Y., Yu, F., and Darrell, T. (2017). End-to-end learning of driving models from large-scale video datasets. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3530–3538.

Zhang, Y. and Yeung, D.-Y. (2010). A convex formulation for learning task relationships in multi-task learning. In *UAI*.