Semantic Segmentation and Depth Estimation of Urban Road Scene Images using Multi-Task Networks

Mohamed Mahyoub
Tutor Reach
United Kingdom
m.mahyoub@tutorreach.com

Friska Natalia
Faculty of Engineering and Informatics
Universitas Multimedia Nusantara
Tangerang, Indonesia
friska.natalia@umn.ac.id

Sud Sudirman
School of Computer Science and Mathematics
Liverpool John Moores University
Liverpool, United Kingdom
s.sudirman@ljmu.ac.uk

Abdulmajeed Hammadi Jasim Al-Jumaily
Universiti Putra Malaysia
Serdang, Malaysia
a.aljumaily@ieee.org

Panos Liatsis
Khalifa University
Abu Dhabi, United Arab Emirates
panos.liatsis@ku.ac.ae

***Abstract*— In autonomous driving, environment perception is an important step in understanding the driving scene. Objects in images captured through a vehicle camera can be detected and classified using semantic segmentation and depth estimation methods. Both these tasks are closely related to each other and this association helps in building a multi-task neural network where a single network is used to generate both views from a given monocular image. This approach gives the flexibility to include multiple related tasks in a single network. It helps reduce multiple independent networks and improve the performance of all related tasks. The main aim of our research presented in this paper is to build a multi-task deep learning network for simultaneous semantic segmentation and depth estimation from monocular images. Two decoder-focused U-Net-based multi-task networks that use a pre-trained Resnet-50 and DenseNet-121 which shared encoder and task-specific decoder networks with Attention Mechanisms are considered. We also employed multi-task optimization strategies such as equal weighting and dynamic weight averaging during the training of the models. The corresponding models' performance is evaluated using mean IoU for semantic segmentation and Root Mean Square Error for depth estimation. From our experiments, we found that the performance of these multi-task networks is on par with the corresponding single-task networks.**

Keywords—Urban Road Scene Analysis; Deep Learning; Multi-Task Networks; Semantic Segmentation; Depth Estimation

I. INTRODUCTION

Vehicle environment perception is an important and preliminary step in understanding the driving scene in an autonomous vehicle. To accurately map the surrounding environment from the vehicle, various range and depth sensors including radar, ultrasonic, and lidar sensors can be used. Camera modules are used to obtain a real-time inference of the objects in a scene [1]. It is of utmost importance to have an accurate and robust perception prediction system since this information is used by safety-critical driver assistance algorithms like cruise control, lane change assists, automatic lane parking, and automatic emergency braking [2]. The main components in scene understanding are object identification and its corresponding depth estimate. This can be achieved by using semantic segmentation and depth estimation of a scene.

Semantic segmentation involves pixel-to-pixel prediction for the corresponding class of object it belongs to in the original image. There are several methods designed to address this issue, including some of the prominent state-of-the-art deep learning methods including the Mask Region-based Convolutional Neural Networks (Mask R-CNN) [3], Recurrent Neural Network-based methods [4], [5], and Encoder-decoder based fully convolutional neural networks [6]. In addition, to improve the accuracy of the prediction, multi-scale and multi-feature extracting encoder-decoder networks were also proposed [7], [8].

Depth estimation involves predicting the relative distance of each pixel in the scene from the viewpoint. It is comparatively a complex and challenging problem because a piece of three-dimensional information has to be inferred from a two-dimensional image space. Depth estimation methods can be broadly categorized into supervised, semi-supervised, and unsupervised-based predictions. In the supervised-based method, the target ground truth depth estimates are available. A convolution neural network-based encoder-decoder network is used to find an absolute depth estimate in [9], [10]. Several graphical-based networks were designed to establish a probabilistic relationship between the neighborhood. One such method is based on the condition random field estimate [11]. This method used the depth estimate to give a scale-invariant prediction that leads to relative-depth-based networks. On the other hand, unsupervised-based methods often use stereo-based image disparity to generate depth estimation of a scene [12] whereas semi-supervised-based methods use auxiliary information from radar, Lidar, and surface normal estimates for depth estimation [13].

Depth estimation and semantic segmentation are closely related tasks and this association has allowed researchers to build a multi-task network for training both tasks using a single network. Thus, a single network can be used to generate both views of an image. It has been empirically shown that multi-task networks can outperform the corresponding single-task networks for a related task [14]. The difference between the single-task and multi-task networks is illustrated in Figure 1.

The main aim of this research project is to build a multi-task learning network for joint semantic segmentation and depth map estimation from a monocular image in an urban driving scene. We will achieve this by investigating suitable data preprocessing steps for the multi-task learning model, building suitable multi-task learning models that are trained using hyperparameters optimization and other multi-task optimization strategies, and by evaluating the performance of the trained multi-task learning model in predicting the semantic segmentation and estimating the depth.

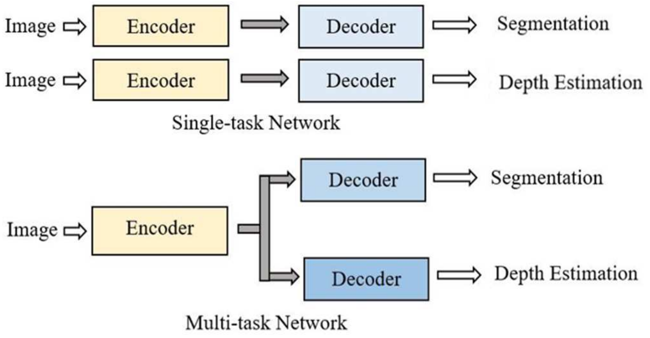

Fig. 1. Architectural difference between single-task and multi-task networks.

II. Literature Review

A. Deep learning-based Semantic segmentation

The prominent semantic segmentation network in deep learning evolved from various kinds of architecture starting from region-based proposal networks like Mask R-CNN [3]. Mask R-CNN network generates object instance segmentation masks on top of the Faster R-CNN region proposal method [15] to obtain pixel-wise segmentation. The path aggregation network [16] is an extension of the mask-R-CNN network that fuses multi-level features into the network for obtaining more accurate predictions. The main advantage of using a region proposal network is that it generates both segmentation and classification of an object in a single network pass. It lacks pixel-to-pixel alignment between input and output and one of the major challenges with this method is to fine-tune the region proposals based on a certain hyperparameter threshold.

A breakthrough in semantic segmentation came with the introduction of encoder-decoder-based network architecture. In this approach, the encoder performs standard feature extraction of the given image and this could be any of the commonly used networks such as VGG [17] and ResNet [3]. The decoder network consists of up-convolution or up-sampling layers with a skip connection from corresponding encoder layers. It utilizes feature information from various layers to build predictions from the coarse level to the much finer pixel level. And, this contributed to an improvement in per-pixel prediction in the case of semantic segmentation.

B. Deep learning-based Depth estimation

Depth estimation from a monocular image is a challenging and ill-posed problem where 3D information has to be reconstructed from a 2D image. Despite this, several deep learning methods have achieved landmark performance in depth prediction. However, the challenge lies in the construction of ground truth labels. The most prominent methods are based on a convolutional neural network with encoder and decoder architecture that was mentioned previously. The encoder extracts a global pool of features using convolution operation which is passed on to the decoder network to fine-tune the final prediction using up-sampling operations. Depth map prediction from a single image using a Multi-Scale Deep Network [9] was the first paper to introduce monocular depth estimation using a convolution network. It consists of a coarse prediction network for extracting global features which are concatenated to the fine-scale network for giving an absolute depth estimate.

Most of the CNN-based supervised methods use absolute depth for training the network. One such method introduced two-streamed parallel networks for estimating fine-scaled depth maps from a single RGB image [10]. One network is for predicting end-to-end depth maps and the other is for obtaining depth gradients of an image that help in fusing structural clues to the network. Both predictions are then fused to get a finer depth estimate. When compared to absolute depth, a relative depth map gives a scale-invariant prediction and produces better generalization. One such method, monocular depth estimation using relative depth maps [18] uses an encoder network for feature extraction and a multiple decoder network for generating relative and ordinary depth maps at multiple scales which are combined to give a final depth estimate. Another method [19] uses a similar approach to generate depth estimates at multiple scales which are then fused using a Condition Random Field model to generate finer depth estimates.

C. Multi-task Learning

As the name suggests, Multi-task Learning (MTL) is an approach to learning multiple related tasks simultaneously in a single network architecture. The inductive transfer mechanism between the tasks helps in better generalization between tasks considered. In many cases, they perform better than their single-task counterparts. MTL works better because of statistical data amplification which increases the sample size of the related task, attribute selection which is a selection of commonly shared feature attributes, eavesdropping that is easing out the learning capability, and representational bias [14]. Learning both semantic and depth maps simultaneously using a multi-task learning framework will help in better generalisability and more accurate prediction than their single-task network since both tasks are related and share common geometrical and structural features.

MTL architecture consists of shared and task-specific layers. The shared layers give a generic representation between all the related tasks and it could be either a hard or soft parameter sharing. Hard parameter sharing has the same hidden representation between the tasks and thus shares common weights between the tasks. Soft parameter sharing is a standalone network for each task where the distance between the tasks is considered for network optimization. Task-specific features are obtained using a task-specific network. MTL architecture is divided into two categories; one with an encoder-focused network and the other with a decoder-focused network. The encoder-focused network allows the task to share parameters at the encoder stage before they are processed with task-specific headers. These networks directly predict the task outputs for a given input in a single processing cycle. Thus, failing to capture commonalities and differences among the results of the tasks resulted in a moderate performance. This is overcome by a decoder-focused network where the network makes an initial prediction of the tasks and later leverages this prediction to refine the final task output.

The prominent encoder-focused model with hard parameter sharing includes Tasks-Constrained Deep Convolutional Network [20] which aims to detect landmark and face alignment tasks. The network supports auxiliary tasks like gender, expression, and appearance attributes. It consists of a shared encoder unit for feature extraction and task-specific header units for predicting individual tasks. This network outperformed the face alignment methods even with occlusions and pose variations.

The encoder-focused soft parameter-sharing network includes Cross-Stitch Networks (CSN) [21] that learn task relatedness by allowing input to each layer as a linear combination of outputs from previous layers. The network architecture consists of two parallel independent networks for each task. These networks are connected by a cross-stitch network for sharing task-related information. Sluice network [22] generalizes the idea of CSN. Each layer is composed of task-specific and shared entries that are orthogonal to each other. Input to the next layer will be a linear combination of these parameters, allowing the network to focus on task-specific or shared values depending on task-relatedness. The Neural Discriminative Dimensionality reduction [23] approach allows for feature fusion at various layers of single-task networks. It uses convolution operation to generate reduced discriminative features acting as an input for task-specific headers. It is similar to CSN where the non-diagonal elements of the weight matrix are set to zero. One of the bottlenecks with this kind of architecture is deciding the sharable and task-specific parameter space. This can lead to suboptimal results and is not a scalable solution.

Along with the above-mentioned networks, other interesting algorithms focused on obtaining the results in real-time [24]. Some networks improved the optimization strategy by dynamically adapting the weights associated with the multi-task loss function for finer prediction [25]. Several other kinds of MTL networks can be referred to through these [26], [27] survey papers.

III. Material and Method

A. Dataset

In this study, we used the images from the Cityscapes [28] dataset. It is an outdoor dataset for urban road scene understanding that consists of high-resolution street-view images of 50 different cities. It is captured over several months in the daytime with good or medium weather conditions. It consists of stereo imagery, with instance and semantic segmentation of 30 object classes categorized as flat, human, vehicle, construction, object, nature, sky, and void type. It also includes a depth map for each scene. The dataset consists of 5000 annotated images with fine annotation. The dataset is split into 2975 training data, 500 validation data, and 1525 testing data. Each image is captured with a resolution of 1024 x 2048. A sample image from the Cityscapes dataset showing the monocular, semantic, and depth map views of the same scenery is shown in Figure 2.

A count plot was constructed to visualize the class distribution in the training dataset. It measures the presence of a class in a given image. The other important parameter to check is the percentage area occupancy of a class. In each image, the percentage area occupancy of a class is measured and an average value is obtained over the entire dataset. These distributions are shown in Figure 3.

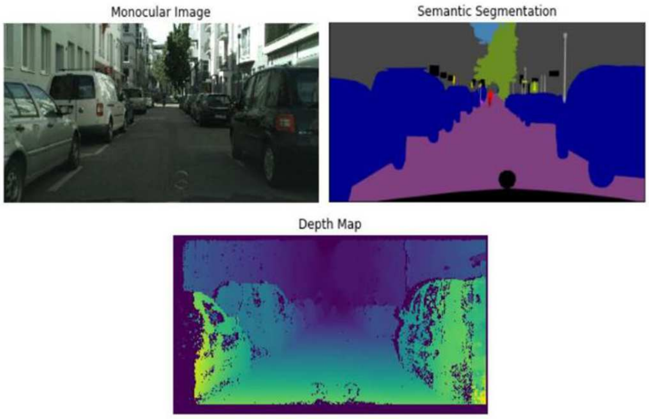

Fig. 2. Sample image from the Cityscapes dataset showing the monocular, semantic segmentation, and depth map view of the same scenery.

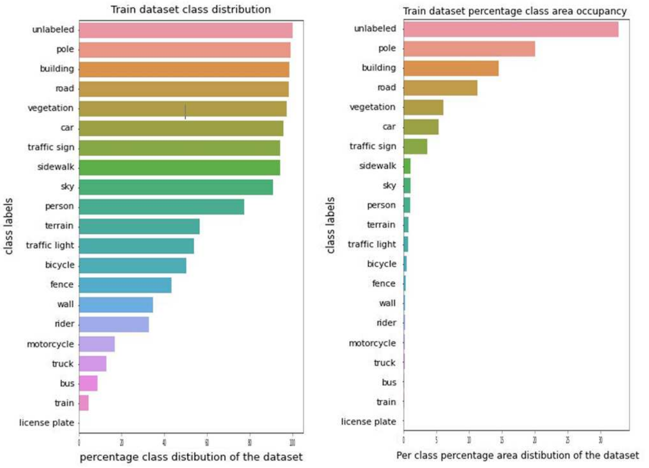

Fig. 3. Percentage Class distribution (left) and per-class percentage area distribution (right) in the training dataset.

B. Semantic Segmentation and Depth Estimation Multi-Task Algorithm

In this study, we will use U-Net-based network architecture with ResNet50 [3] and DenseNet121 [29] as the backbone. The U-Net-based network architecture consists of a contracting path and an expanding path. The contracting path is used for feature abstraction from low-level information to a more compact representation and the expanding path is used for the precise localization of objects. The contracting path consists of repeated units of convolution layers followed by Rectified Linear Unit activation function. Features are down-sampled using the max-pooling operation. The expansion path consists of an up-sampling unit for expanding the feature map. It is concatenated with the corresponding cropped feature map from the contracting path to obtain a more precise output representation [30]. MTL architecture can use a contracting path as a shared unit between tasks and an expanding path for a task-specific decoder unit [31]. The single-task network architecture of the U-Net models using ResNet50 and DenseNet121 are shown in Figures 4 and 5, respectively.

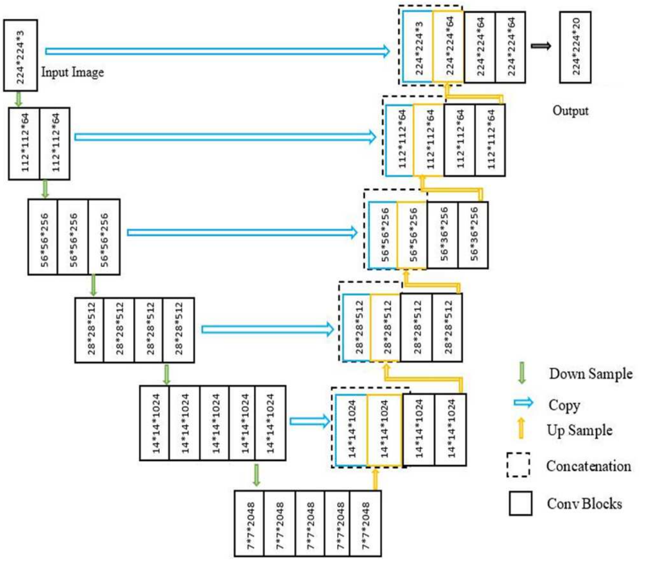

Fig. 4. U-Net with ResNet-50 backbone for either semantic segmentation or depth estimation task.

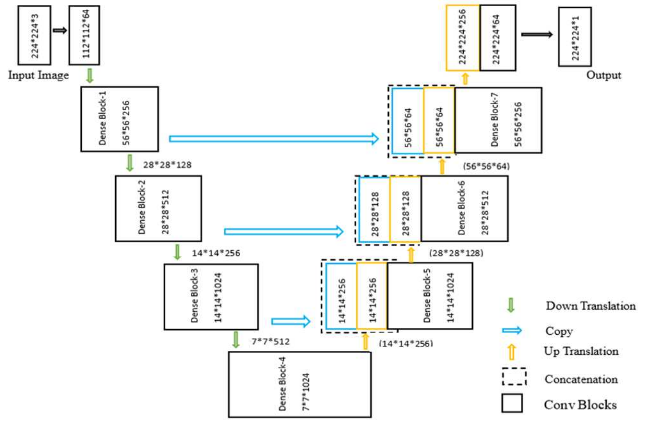

Fig. 5. U-Net with DenseNet-121 backbone for either semantic segmentation or depth estimation task.

We adopted attention networks to improve the fusing of data in the decoder. It enables the passing of task-correlative information to the individual task-specific network to fine-tune the prediction [32]. Attention-based network mimics the human cognition system, that is the ability to focus on relevant objects and to ignore non-relevant objects that are present in a given scene. These networks gained more popularity because of their ability to focus on the important features that are beneficial for a given task by a gating mechanism. Attention mechanisms can be integrated into the MTL network for obtaining spatial or temporal attention [33]–[35]. For semantic segmentation and depth estimation tasks, attention networks can help in focusing the contextual information present in a scene and help in improving the prediction accuracy [19], [31], [36]. The multi-task network architecture of the U-Net models using ResNet50 and DenseNet121 are shown in Figures 6 and 7, respectively.

Training a multi-task network involves selecting a suitable network architecture and a corresponding loss function for each task. The overall multi-task loss function can be optimized using Stochastic Gradient Descent (SGD) with momentum or ADAM-based optimization algorithms [37]. In the case of SGD optimization, a suitable learning rate and momentum have to be selected. In the case of ADAM optimization, a suitable initial learning rate has to be selected. The exact optimization algorithm used in the experiment is described in the next section.

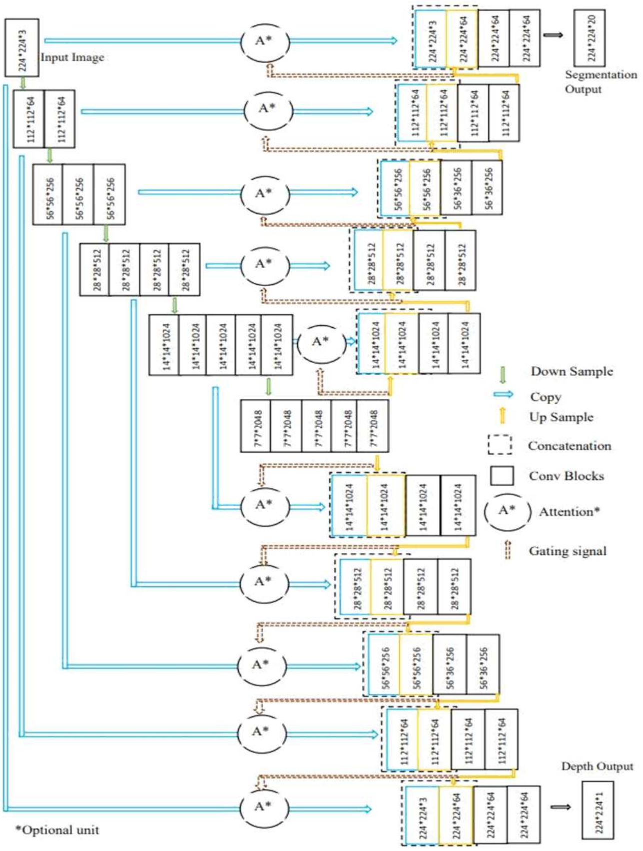

Fig. 6. U-Net with ResNet-50 backbone for both semantic segmentation and depth estimation tasks.

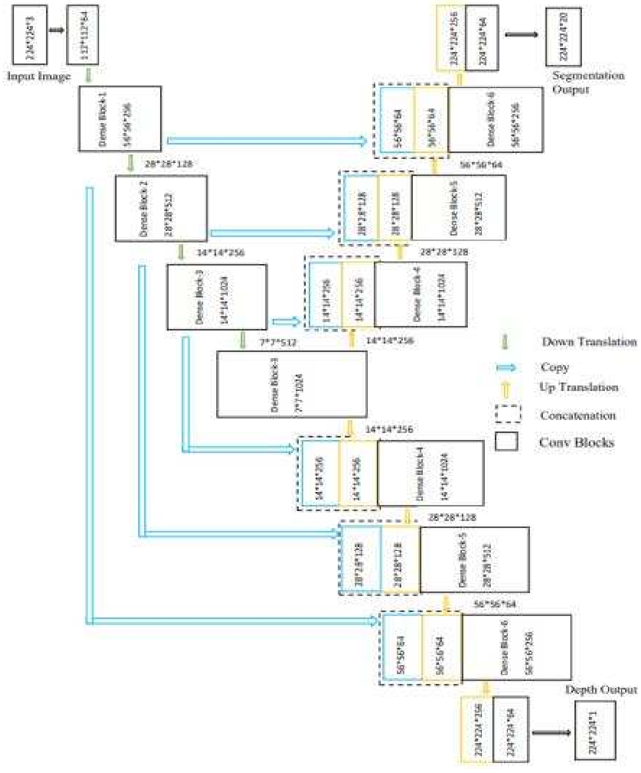

Fig. 7. U-Net with DenseNet-121 backbone for both semantic segmentation and depth estimation tasks.

IV. Experiment and Result Analysis

As described previously, we used ResNet50 and DenseNet121 network architecture for both the single-task and multi-task models. The overall training data is split into mini-batches containing 2, 4, or 8 images per batch. Depending on the batch size the total number of epochs is adjusted which starts from 50 and can go up to 200. This can also be parameterized depending on the gradient update or learning rate per epoch.

Two loss weighting strategies were adopted in our experiment, they are Gradient Weighting and Geometric Loss (GWGL) and Equal Weighting and Dynamic Weight Averaging (EWDW) [38]. Therefore, in total, we have four multi-task models – in addition to the four single-task models. All models are trained to use ADAM optimizer with an initial learning rate of 0.0001. A polynomial learning rate decay of 0.9 is used over the epochs with a min-batch size of 8. The model is trained with cross-entropy loss and a dice loss individually. All four single-task models are trained with an epoch search range between 70 and 85 epochs whereas both multi-task models have been trained with an epoch search range between 80 and 100 epochs.

The mean IOU score is used as an evaluation metric for semantic segmentation and Root Mean Square Error for depth estimation. The results are shown in Table I.

TABLE I. Models Performance in Semantic Segmentation (Mean IoU) and Depth Estimation (RMSE)

		RMSE	**Mean IOU**
Single-Task	ResNet50	**0.042**	34.76
	DenseNet121	0.043	**36.37**
Multi-Task	ResNet50 EWDW	0.050	**33.73**
	DenseNet121 EWDW	**0.043**	33.64
	ResNet50 GWGL	0.045	31.21
	DenseNet121 GWGL	0.950	30.01

The result shows that multi-task models trained using Equal Weighting and Dynamic Weight Averaging strategy produce a comparable performance to the single-task counterparts. On the other hand, multi-task models trained using Gradient Weighting and Geometric Loss strategy seem to lag in terms of performance, and more over seem to have failed miserably when performing the depth estimation task.

V. Conclusion

In this paper, we have presented our work in investigating the performance of a multi-task network approach to semantic segmentation and depth estimation of urban road scene images. We developed eight models (four single-task and four multi-task) using U-Net architecture with either ResNet50 or DenseNet121 network as the backbone. The multi-task models are trained using two loss weighting strategies namely the Gradient Weighting and Geometric Loss and Equal Weighting and Dynamic Weight Averaging strategies. We found that the multi-task models trained using Equal Weighting and Dynamic Weight Averaging strategy (regardless of backbone architecture choice) produce a comparable performance to the single-task models. The framework that we have developed can also be extended to other network topologies and optimization strategies. This gives the flexibility to include more related tasks with a shared encoder and a task-specific decoder configuration.

References

[1] B. Wang, Y. Han, D. Tian, and T. Guan, “Sensor-based environmental perception Technology for Intelligent Vehicles,” *J. Sensors*, vol. 2021, 2021.

[2] A. Paul, R. Chauhan, R. Srivastava, and M. Baruah, “Advanced Driver Assistance Systems,” *SAE Tech. Pap.*, vol. 2016-Febru, no. February, 2016.

[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[4] B. Shuai, Z. Zuo, B. Wang, and G. Wang, “Dag-recurrent neural networks for scene labeling,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3620–3629.

[5] A. Salvador, M. Bellver, V. Campos, M. Baradad, F. Marques, J. Torres, and X. Giro-i-Nieto, “Recurrent neural networks for semantic instance segmentation,” *arXiv Prepr. arXiv1712.00617*, 2017.

[6] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation,” *CoRR*, vol. abs/1706.0, 2017.

[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[9] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[10] J. Li, R. Klein, and A. 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. 3372–3380.

[11] D. Xu, W. Ouyang, X. Wang, and N. Sebe, “Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 675–684.

[12] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.

[13] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, “Geonet: Geometric neural network for joint depth and surface normal estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 283–291.

[14] R. Caruana, “Multitask learning,” *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[15] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[16] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[17] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in *International Conference on Learning Representations*, 2015.

[18] J.-H. Lee and C.-S. Kim, “Monocular depth estimation using relative depth maps,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9729–9738.

[19] E. Ricci, W. Ouyang, X. Wang, N. Sebe, and others, “Monocular depth estimation using multi-scale continuous CRFs as sequential deep networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1426–1440, 2018.

[20] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, “Learning deep representation for face alignment with auxiliary attributes,” *IEEE*

Trans. Pattern Anal. Mach. Intell., vol. 38, no. 5, pp. 918–930, 2015.

[21] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, “Cross-stitch networks for multi-task learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3994–4003.

[22] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard, “Latent multi-task architecture learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 01, pp. 4822–4829.

[23] Y. Gao, J. Ma, M. Zhao, W. Liu, and A. L. Yuille, “NDDR-CNN: Layerwise feature fusing in multi-task CNNs by neural discriminative dimensionality reduction,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3205–3214.

[24] V. Nekrasov, T. Dharmasiri, A. Spek, T. Drummond, C. Shen, and I. Reid, “Real-time joint semantic segmentation and depth estimation using asymmetric annotations,” in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 7101–7107.

[25] A. Kendall, Y. Gal, and R. Cipolla, “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*, 2018, pp. 7482–7491.

[26] M. Crawshaw, “Multi-task learning with deep neural networks: A survey,” *arXiv Prepr. arXiv2009.09796*, 2020.

[27] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, “Multi-task learning for dense prediction tasks: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[28] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[30] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9351.

[31] A. Jha, A. Kumar, B. Banerjee, and S. Chaudhuri, “Adamt-net: An adaptive weight learning based multi-task learning model for scene understanding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 706–707.

[32] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, “Pattern-affinitive propagation across depth, surface normal and semantic segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4106–4115.

[33] S. M. R. Hashemi, “A survey of visual attention models,” *Ciência e Nat.*, vol. 37, pp. 297–306, 2015.

[34] K. K. Evans, T. S. Horowitz, P. Howe, R. Pedersini, E. Reijnen, Y. Pinto, Y. Kuzmova, and J. M. Wolfe, “Visual attention,” *Wiley Interdiscip. Rev. Cogn. Sci.*, vol. 2, no. 5, pp. 503–514, 2011.

[35] V. Mnih, N. Heess, A. Graves, and others, “Recurrent models of visual attention,” *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[36] S. Liu, E. Johns, and A. J. Davison, “End-to-end multi-task learning with attention,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1871–1880.

[37] S. Ruder, “An overview of gradient descent optimization algorithms,” *CoRR*, vol. abs/1609.0, 2016.

[38] T. Gong, T. Lee, C. Stephenson, V. Renduchintala, S. Padhy, A. Ndirango, G. Keskin, and O. H. Elibol, “A comparison of loss weighting strategies for multi task learning in deep neural networks,” *IEEE Access*, vol. 7, pp. 141627–141632, 2019.