Semantic Segmentation for Self-Driving Cars

Ahmed Hesham Aboualy

Hussin Mohamed Elrashidy

Abdelrahman Muhsen Abdelghany

Mariam Magued Bebawy

Abstract*— Self-driving cars are considered a work-in-progress as they're still being developed by companies such as Tesla and Nvidia. Although they are making more complicated systems that may include sensors, lidars, and high-quality cameras, in this paper, we implement a semantic segmentation model, using a deep learning approach, for images taken from real streets and cluster the important elements in said images; cars, pedestrians, streets, motorcycles, walls, and fences, so that a car will be able to move through the segmented output. This paper presents a novel approach to semantic segmentation using HRNet and Transfer Learning. Our model takes advantage of the HRNet architecture, which provides a high-resolution representation of the input image, to accurately localize and classify objects in the scene. We further leverage Transfer Learning to initialize our model weights with pre-trained weights on a large-scale dataset, which are then fine-tuned on the target dataset. Our experiments demonstrate that our proposed model is comparable to the state-of-the-art semantic segmentation methods on several datasets [1], given the computational efficiency and accuracy obtained through limited resources.

Keywords—Semantic Segmentation, HRNet, Transfer Learning, Deep Learning, Self-driving Cars

I. Introduction

Semantic segmentation is an important task in image analysis and computer vision, which involves classifying each pixel in an image into a specific class. Traditional methods, such as thresholding and histogram-based segmentation, have proven to be effective in some cases, but they have limited accuracy and lack the ability to generalize to different data types or applications. In recent years, deep learning-based methods have become the state-of-the-art for semantic segmentation.

Semantic segmentation through a deep learning technique is used to assign labels to individual pixels in an image. It is a pixel-wise classification task that can be used to segment objects from a background in an image, which can be deployed to use in a variety of applications such as autonomous driving and medical imaging. CNNs are often used, and HRNet is a popular choice for this specific task due to its superior performance. Transfer learning is also commonly used, which involves taking a pretrained model and fine-tuning it on a new dataset. This can greatly reduce the time and cost required to train a model, as well as improve accuracy. Our technique typically involves designing a custom HRNet suitable for our obtained dataset, which is then combined with a pretrained model in order to obtain better results.

a deep convolutional neural network that is trained on a large dataset of labeled images [3]. This deep network can then be used to predict the class of each pixel in an image, allowing for accurate segmentation. Additionally, semantic segmentation can be used to associate each pixel of the image with a class [3], allowing for a more detailed understanding of the content of the image. This can be used for a variety of tasks such as recognizing objects, detecting edges, and identifying features.

II. Related Work

(Li et. al., 2018) [4] resort to transfer knowledge from automatically rendered scene annotations in virtual world to facilitate real-world visual tasks. (Yuan et. al., 2019) [5] address the semantic segmentation problem with a focus on the context aggregation strategy. (Bai et. al., 2020) [6] propose a novel deep neural network architecture, named MPDNet, for fast and efficient semantic segmentation under resource constraints due to a large number of parameters and floating point operations. To overcome the problem (Zheng et. al., 2020) [7] propose explicitly estimating the prediction uncertainty during training to rectify the pseudo label learning for unsupervised semantic segmentation adaptation, Given the input image, the model outputs the semantic segmentation prediction as well as the uncertainty of the prediction. (Wang et. al., 2020) [8] propose a simple and flexible two-stream framework named Dual Super-Resolution Learning (DSRL) to effectively improve segmentation accuracy without introducing extra computation costs. (Yuan et. al., 2020) [9] address the semantic segmentation problem with a focus on the context aggregation strategy, their architecture was similar to our approach (HRNET) but in addition to object-contextual representation (OCR). (Huang et. al., 2020) [10] focus on a novel topic, weakly-supervised semantic segmentation in cityscape via HSIs. (Nirkin et. al., 2021) [11] present a novel, real-time, semantic segmentation network in which the encoder both encodes and generates the parameters (weights) of the decoder. They made a new type of hyper network, composed of a nested U-Net for drawing higher level context features, a multi-headed weight-generating module that generates the weights of each block in the decoder immediately before they are consumed, for efficient memory utilization, and a primary network that is composed of novel dynamic patch-wise convolutions. To improve reliability (Franchi et. al., 2021) [12] introduce Superpixel-mix, a new superpixel-based data augmentation method with teacher-student consistency training. (Liu et. al., 2022) [13] propose a simple yet effective feature distillation method called normalized feature distillation (NFD), aiming to enable effective distillation with the original features without the

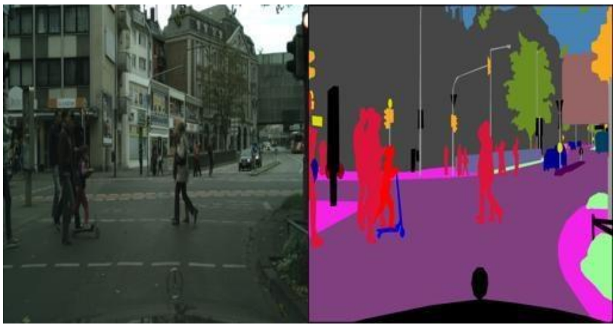

fig 1; sample image from CityScape dataset need to manually design new forms of knowledge.

III. Dataset & Features

The Cityscapes Dataset is a large-scale dataset which focuses on semantic understanding of urban street scenes. It contains high-quality stereo video sequences recorded in 50 different cities from around the world[14][15], as well as dense pixel-level, instance-level, and panoptic annotations. The dataset is used to evaluate the performance of vision algorithms for major tasks of semantic urban scene understanding, such as object detection, autonomous navigation, and scene understanding. It can also be used to train algorithms for autonomous driving and other applications. The dataset includes images of various resolutions, as well as labels such as street scenes, vehicles, pedestrians, and lane markings. Algorithms; such as those for autonomous navigation, can be developed through the help of this dataset to improve the safety of autonomous vehicles.

Due to computational reasons and limited resources, we use only a subset of the dataset of a total number of 3475 images and 34 segmented labels, an example is shown in *fig-1*. Preprocessing is an important step when working with this particular dataset. This includes the splitting of images to image and mask pairs, resizing of images to 128 x 128 pixels, and normalization of values. Additionally, it is important to ensure that the data is split into train, validation, and test sets in order to avoid overfitting. Training, validation sets are split as follows 2975, and 500 images.

IV. Methods

The approach we are presenting for semantic segmentation is High-Resolution Net (specifically HRNetV2), which can keep high-resolution representations intact throughout the process. We begin with a high-resolution convolution stream, then add high-to-low-resolution convolution streams one by one before connecting the multi-resolution streams in parallel. As shown in *fig-2* (by Wang et al., 2019), the final network is comprised of numerous (4 in this paper) stages, with the nth stage containing n streams corresponding to n resolutions. We perform repeated multi-resolution fusions by exchanging data across parallel streams.

HRNet's high-resolution representations are not only semantically powerful but also spatially exact. This is due to two factors. Rather than connecting high-to-low-resolution convolution streams in series, this approach joins them in parallel. As a result of this approach's ability to preserve high resolution rather than recovering high resolution from low resolution, the learned representation may be spatially more exact. The majority of extant fusion approaches combine high-resolution low-level and high-level representations derived from upsampling low-resolution representations. Instead, they perform multiresolution fusions to increase high-resolution representations using low-resolution representations, and vice versa. As a result, all of the high-to-low-resolution representations have strong semantics.

The final stage is the primary distinction between the three types of HRNet. In HRNetV1, the only output is from the high-resolution convolution stream, whereas in HRNetV2 (the one we propose), the final stage consists of the concatenation of different resolutions (the high-resolution with the up-sampling of three others), and in HRNetV2p, the same as in HRNetV2 but with the HRNetV2 feature map pyramid.

In our case, and due to the lack of data, we simplified the architecture proposed by (Wang et al., 2019) to consist of only 3 stages of convolution; bottleneck – two parallel streams with 1/4, 1/8 resolutions – three parallel streams with 1/4, 1/8, 1/16 resolutions, as shown in *fig-3*.

V. Results & Discussion

In order to start implementing our model, some preprocessing was done for the data. The number of classes was reduced from 34 to 9 classes for easier data handling. Moreover, the segmented masks were color encoded so as to be fed as the output of the training variables.

Initially, the implementation was focused on the HRNet architecture solely. The implementation utilized most of the basic blocks of this network in order to obtain a model

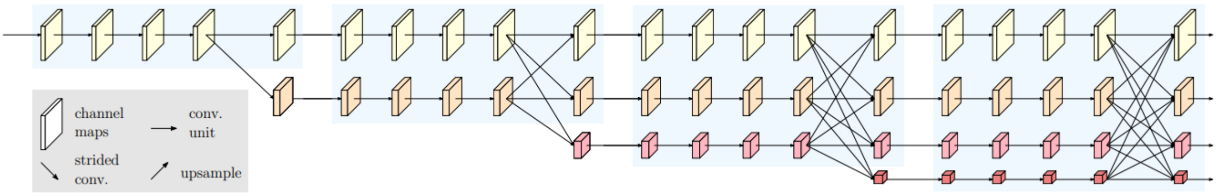

fig 2; architecture presented by Wang et al.

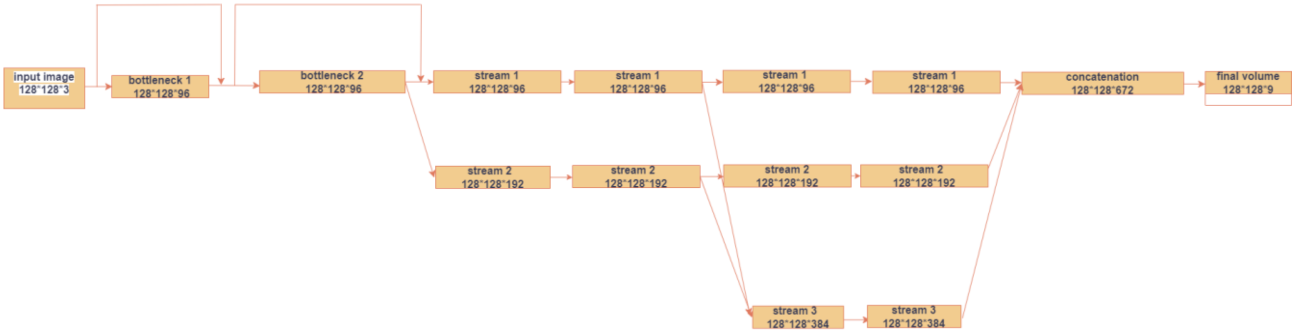

fig 3; model architecture

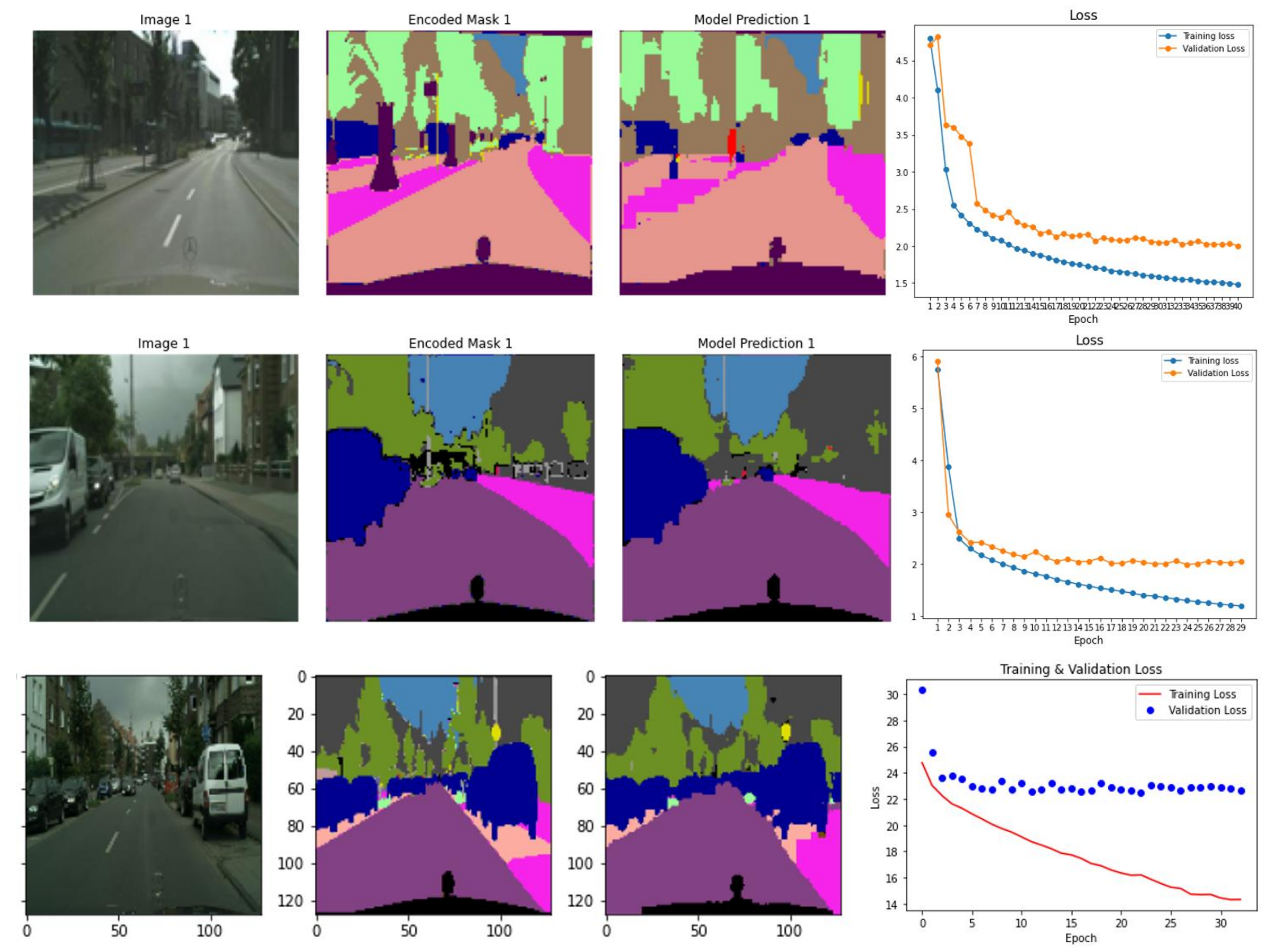

fig 4; model predictions and loss; a) 1st model; b) 2nd model; c) 3rd model

suitable to the resources and task at hand. The model consisted of roughly 14 million parameters. This model managed to reach an accuracy of 84.5% and a mean IoU of 61.5%.

The following implementation involved utilizing the techniques of transfer learning in order to make use of a pretrained model –namely MobileNetV2, which is trained on object classification on the dataset ImageNet– to enhance the predictions of our model. 4 blocks of this network was used and concatenated with our HRNet-based model, therefore reaching ~28 million parameters. This model showed minor enhancements in terms of the evaluation metrics; accuracy of 85% and mean IoU of 62.8%.

The final implementation didn't require any additional preprocessing; the number of classes was kept at 34 as the original dataset. This model also utilized transfer learning; it was basically the same model as before, just enhanced to be able to train on all 34 classes as intended. This model's parameters were about 28 million as well, and managed to reach 80.6% accuracy but only 27.1% for the mean IoU. To explain, as we increase the number of classes, the number of misclassifications between said classes increase. That may be due to multiple classes being close to each other in terms of the features extracted, e.g., "ground" and "sidewalk". While the mean IoU may be low, this model portrayed the best results when it comes to segmenting the "road" class, which is ultimately one of the most important classes for a self-driving car.

VI. Conclusion & Future Work

In conclusion, semantic segmentation using HRNet and transfer learning proved to be effective methods for performing computer vision tasks such as object detection, autonomous navigation, and scene understanding. HRNet is a powerful architecture which has superior performance when compared to other models, and transfer learning can reduce the time and cost required to train a model, as well as improve accuracy. By combining these two techniques, it is possible to quickly and effectively develop algorithms for various tasks such as this one.

In the future, through the use of more resources, further improvement to the performance of our model can be achieved by increasing the spatial resolution of the system inputs. One potential area of research could be to explore more efficient architectures and methods to reduce the computational cost of the model while maintaining accuracy. Another potential area of research could be to explore methods to incorporate recurrent neural networks (RNN) to improve prediction. Finally, research could also be done to explore methods to incorporate domain knowledge into the model, which could improve the model's ability to generalize to new data.

References

[1] Wang, X., Xie, Y., & Chen, Y. (2020). Dual Super-Resolution Learning for Semantic Segmentation. Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition, 3684-3693. https://openaccess.thecvf.com/content_CVPR_2020/papers/Wang_Dual_Super-Resolution_Learning_for_Semantic_Segmentation_CVPR_2020_paper.pdf

[2] Joshi, R. (2020). Introduction to Semantic Image Segmentation. Analytics Vidhya. https://medium.com/analytics-vidhya/introduction-to-semantic-image-segmentation-856cda5e5de8

[3] MathWorks. (n.d.). Image segmentation. Retrieved January 10, 2023, from https://www.mathworks.com/discovery/image-segmentation.html

[4] Peilun Li; Xiaodan Liang; Daoyuan Jia; Eric P. Xing; (2018). "Semantic-aware Grad-GAN For Virtual-to-Real Urban Scene Adaption", ARXIV

[5] Yuhui Yuan; Xiaokang Chen; Xilin Chen; Jingdong Wang; (2019). "Segmentation Transformer: Object-Contextual Representations for Semantic Segmentation" , ARXIV

[6] Xing Bai; Jun Zhou; "Efficient Semantic Segmentation Using Multi-Path Decoder", APPLIED SCIENCES

[7] Zhedong Zheng; Yi Yang; (2020). "Rectifying Pseudo Label Learning Via Uncertainty Estimation For Domain Adaptive Semantic Segmentation", ARXIV

[8] Li Wang; Dong Li; Yousong Zhu; Lu Tian; Yi Shan; (2020). "Dual Super-Resolution Learning for Semantic Segmentation", CVPR

[9] Yuhui Yuan; Xilin Chen; Jingdong Wang; (2020). "Object-Contextual Representations For Semantic Segmentation", ECCV

[10] Yuxing Huang; Shaodi You; Ying Fu; Qiu Shen; (2020). "Weakly-supervised Semantic Segmentation in Cityscape Via Hyperspectral Image", ARXIV

[11] Yuval Nirkin; Lior Wolf; Tal Hassner; (2021). "HyperSeg: Patch-Wise Hypernetwork for Real-Time Semantic Segmentation", CVPR

[12] Gianni Franchi; Nacim Belkhir; Mai Lan Ha; Yufei Hu; Andrei Bursuc; Volker Blanz; Angela Yao; (2021). "Robust Semantic Segmentation with Superpixel-Mix", ARXIV

[13] Tao Liu; Xi Yang; Chenshu Chen; (2022). "Normalized Feature Distillation for Semantic Segmentation", ARXIV

[14] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes Dataset. Retrieved from https://www.cityscapes-dataset.com/

[15] Wikipedia contributors. (n.d.). List of datasets for machine-learning research. In Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

[16] TensorFlow. (n.d.). TensorFlow. Retrieved from https://www.tensorflow.org/

[17] Chollet, F. (n.d.). Keras: The Python Deep Learning library. Retrieved from https://keras.io/.

[18] TensorFlow. (n.d.). MobileNetV2: Inverted Residuals and Linear Bottlenecks. Retrieved from https://www.tensorflow.org/api_docs/python/tf/keras/applications/mobilenet_v2/MobileNetV2.

[19] Harris, S. (2020). Array programming with NumPy. Nature, 585, 357362. https://doi.org/10.1038/s41586-020-2649-2

SUPPLEMENTARY MATERIAL

[1] 1st model: https://www.kaggle.com/code/ahmedheshamsbe/hr-9classes-diceloss

[2] 2nd model: https://www.kaggle.com/code/sadguava/semseg-tests

[3] 3rd model: https://www.kaggle.com/code/sadguava/hr-net-cityscapes

CONTRIBUTION

[1] *Ahmed Aboualy:*

Experimenting with HRNet architecture
Experimenting with loss functions

[2] *Hussin Elrashidy:*

Initial model architecture
Experimenting with small models

[3] *Abdelrahman Abdelghany:*

Transfer learning modelling
Experimenting with preprocessing

[4] *Mariam Bebawy:*

Experimenting with preprocessing
Experimenting with data augmentation