- **Topic: Image/Video Object Detection**

Paper #1:

Qiao, D., & Zulkernine, F. (2021). *Drivable area detection using Deep Learning models for autonomous driving*. IEEE Xplore. Retrieved September 14, 2022, from https://ieeexplore.ieee.org/document/9671392

Summary: Developing fully autonomous vehicles is a topic that has had a lot of attention in the media in recent years. Having a car be fully autonomous will require several reliable sub-systems, one of which would be being able to detect the drivable area. The drivable area is the area that the car itself should drive in and differs from adjacent lanes of traffic for example. The network should be able to detect whether the adjacent lanes are available to be used in emergency situations, or to pass other vehicles, but should know to stay in its lane under normal conditions. To further complicate this, not all roads have lane markings or consistent markings, but the network still needs to be confident in predictions under these circumstances. All this needs to be accomplished in real-time as the car is travelling.

This paper uses several datasets to train and test the system. It uses the Cityscapes street scene dataset to train the model to differentiate between road and other parts of transportation infrastructure. It also combines this with the BDD100K dataset that includes roadways, that are segmented based on driving lanes, among other features. The model consists of three main components:

1) ImageNet pretrained ResNet backbone
2) Atrous Spatial Pyramid Pooling (ASPP)
3) Feature Pyramid Network

The ResNet is applied to extract deep features from the input traffic images, which is then passed to the ASPP to get rich semantic information and finally the FPN collects rich feature information. The ResNet model has 101 convolutional layers, and 5 residual blocks that are encoders B0, B1, B2, B3, B4. The different blocks will apply different down scaling factors of 2, 4, 8, 16, 32 for encoders B0-B4 respectively. The ResNet has an input resolution of 720x1280 pixels. After the B4 encoder, the ASPP module fuses five different feature maps together to extract contextual features. The outputs of the encoder blocks B1-B4 are transferred to the corresponding decoder blocks F1-F4. The outputs are then scaled back to the original input size and features are merged with encoders outputs.

Some of the limitations are that the detection of drivable area cannot identify potential risks such as if vision is obstructed by other obstacles. Further research is also needed to be able to make use of the drivable area detection to form a trajectory for the car.

Paper #2:

Han, C., Zhao, Q., Zhang, S., Chen, Y., Zhang, Z., & Yuan, J. (2022, August 24). *Yolopv2: Better, faster, stronger for panoptic driving perception*. arXiv.org. Retrieved September 15, 2022, from https://arxiv.org/abs/2208.11434

Summary:
Main contributions are improvements to the YOLO architecture that are summarized as follows:

Better:
- More effective model structure
- Developed more sophisticated bag-of-freebies
- Mosaic and Mixup were performed for data preprocessing

Faster:
- More efficient model structure and memory allocation

Stronger:
- Well generalized for adapting to various scenarios while ensuring speed

Model uses a very similar design concept to YOLOP and HybridNet architecture but makes uses of a powerful backbone for the feature extraction. This model uses three decoder branches to run drivable area segmentation and lane detection tasks, instead of running in one branch. The network architecture makes use of one shared encoder for feature extraction from the input image, that then feeds the three separate decoder heads. Compared to YOLOP which uses CSPdarknet, this model adopts E-ELAN to make use of group convolution. A Path Aggregation Network (PAN) combines features with FPN to fuse semantic information for local features.

Drivable area segmentation and lane segmentation are performed by separate task heads, each with their own structure. The authors found that the features extracted from deeper layers as in YOLOP, were unnecessary for drivable area segmentation. These features do not help performance but increases the difficulty of model convergence during training. The head responsible for drivable area segmentation is connected before the FPN module.

The paper proposed an effective and efficient end-to-end multi-task learning network that simultaneously performs object detection, drivable area segmentation and lane detection. It achieves new state-of-the-art (SOTA) performance on the BDD100K dataset, exceeding previous models in both speed and accuracy.

The model used a Cosine Annealing policy to adjust the learning rate during training. Initially, the learning rate is set to 0.01 and a warm-restart is used in the first 3 epochs. Total training epochs is 300.

Paper #3:

He, K., Zhang, X., Ren, S., & Sun, J. (2015, April 23). *Spatial pyramid pooling in deep convolutional networks for visual recognition*. arXiv.org. Retrieved September 17, 2022, from https://arxiv.org/abs/1406.4729

Summary:
This is a foundational paper in the field of object recognition and image classification and has been cited nearly 9000 times since it was published in 2015. CNNs are very good at classifying images

however they require a fixed image size to work. This limits not only the scale of the image, but also the aspect ratio. Typically, images are either cropped or warped to make it fit the size needed for the network, which can compromise accuracy. CNNs are made up of two parts: 1) convolutional layers and 2) fully-connected layers. The fully-connected layers are the reason why the CNNs require fixed input size. This paper proposes the idea of creating a Spatial Pyramid Pooling (SPP) layer after the convolutional layer to remove the need for fixed image input size, since the SPP will output a fixed size for the fully-connected layers.

The SPP pools features deeper in the network, to remove the need to crop and warp images at the input. SPP-net also allows for the model to be trained on images or varying sizes and scales. This increases the scale-invariance and reduces over-fitting. SPPs have long been recognized as one of the most successful techniques in computer vision. However, until this paper SPP has not been considered in conjunction with CNNs. SPP has various properties that pair well with CNNs.

1) SPP generates a fixed-length output regardless of input size
2) SPP works well with objects that have been deformed
3) SPP works with objects of different scales, due to the way it extracts features

SPP is versatile enough that it can handle different sizes, aspect ratios and scales. Combining SPP with deep neural networks results in SPP-net which has shown incredible results in classification/detection tasks. The authors trained the model using two fixed-length inputs. They had a network for 224x224 images and another for 180x180. The 224x224 images were resized to fit the 180x180 size. This multi-size training was done to simulate the varying input sizes the model needs to learn, while still taking advantage of the well-optimized fixed-size existing implementations. The authors believe that this model was the first of its kind to train a single network on varying sized images.

Implementation is based upon the publicly available code *cuda-convnet* and *Caffe*. This paper used existing network architectures from other publications and modified them to use SPP. They then compared the results to the baseline tests of the architectures without any modifications. The results were overwhelmingly positive for the SPP layer in a CNN architecture.

Comparison Table:

Paper	**Problem Addressed**	**Data Used**	**Models Implemented**	**Results**	**Shortcomings**
Qiao, D., Zulkernine, F. [2021]	Drivable Area Detection	- BDD100K - Cityscapes street scene	- ResNet - ASPP - FPN	Cityscape - 98.03% Pixel Accuracy (PA) - 98.00% mean Pixel Accuracy (mPA) - 95.00%	- Does not work if the vision is obstructed

				mean Intersection over Union (mIoU) BDD100K - 97.09 % PA - 91.14 % mPA - 84.58 % mIoU	
Han, C., et al. (2022)	Drivable Area Detection Lane Detection	- BDD100K	- YOLOP - BoF - HybridNet - FPN - Spatial Pyramid Pooling (SPP) - PAN	BDD100K - 93.00 % mIoU - 0.83 MAP - 87.3% accuracy for lane detection	- Few shortcomings, this paper proposed a new SOTA model.
He, K., et al. (2015)	Object Detection Image Classification	- Pascal VPC 2007 - Caltecth101	- SSP - SSP-net - CNN	This paper tests the idea of implementing SPP with many different types of network architectures. Essentially all of the metrics included in the study suggest that using the SPP was beneficial. The team entered their own network into the Large Scale Visual Recognition Challenge (ILSVRC) 2014 and placed 2nd in object detection and 3rd in image classification, against 38 teams. There are no notable shortcomings from this study.	

Sample Data

Cityscapes

- Includes multiple classes of annotations including road, person, car and vegetation.
- Only the road class was used in the first paper, as it focused on detecting drivable area.

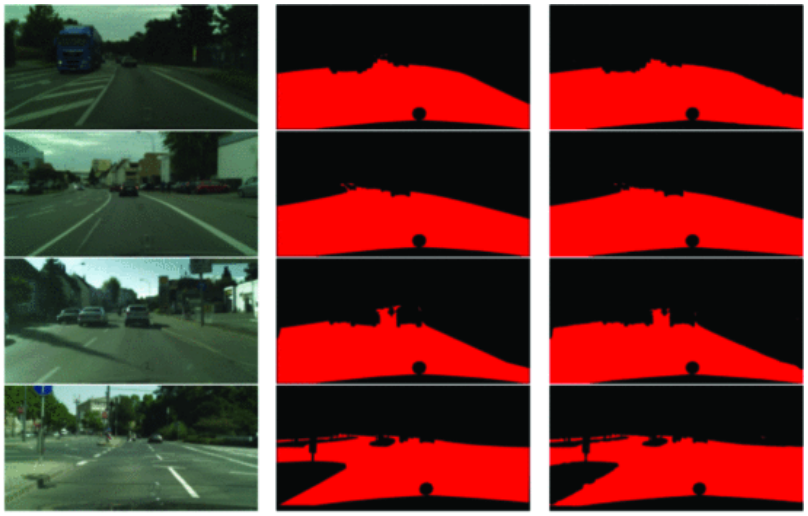

Figure 1 Sample road segmentation data from Cityscapes. a) is the original driving scene image. b) is the annotated data included in the dataset. c) is the model output results

BDD100K

- Distinguishes between direct drivable area, and alternative drivable regions.
- Cityscapes only distinguishes between road and other parts of the image, this dataset further divides the road area

Dataset Name	Size	Type of Data and Features (columns)	# of Training/Testing samples
Cityscapes	5000 Total Images (only a subset was used by Qiao, D. et al)	- 30 classes of annotations, including sidewalk, car, bus, traffic light, person, sky, etc.	2750 Training 500 Validation 1525 Testing
BDD100K	100,000 High resolution (720x1280 pixels) images	- Includes diverse scene types, including city streets, tunnels, residential areas, parking, highways - Multiple weather conditions including clear, overcast, snowy, rainy, cloudy, foggy - Different times of day	70, 000 training 10, 000 validation 20, 000 testing
Caltech101	9144 images	- There are images of 101 (plus a clutter class) object classes. These images in the classes range from saxophones to beavers to roller skates, as an example.	102 categories of images From each category, 30 for training and up to 50 for testing.