**Literature Review**

Pedestrian detection stands as a foundational challenge for researchers aiming to study deep learning algorithms and develop safer forms of travel. Zhang et al. introduced CityPersons as a new set of high-quality annotations aimed to provide a dataset to adapt the FasterRCNN architecture to achieve cutting-edge results on established benchmarks like Caltech and KITTI. The CityPersons dataset was created to provide high-quality bounding box annotations for pedestrians in urban environments, aiming to enhance pedestrian detection research by offering diverse and challenging training data for improved model performance and generalization across multiple benchmarks. The study demonstrated that by adapting the FasterRCNN architecture and training it on the CityPersons dataset, significant improvements were achieved in pedestrian detection performance. The adapted FasterRCNN model showcased state-of-the-art results on established benchmarks like Caltech and KITTI, particularly excelling in detecting small-scale pedestrians and handling heavy occlusions. The FasterRCNN detector, known for its competitive performance in general object detection had underperformed in pedestrian detection tasks on the Caltech dataset due to its inability to detect small-scale objects (50 ∼ 70 pixels), which are prevalent in the dataset. The adapted FasterRCNN model demonstrated superior performance in pedestrian detection tasks, achieving a mean MR of 10.27 on the Caltech dataset and 12.81 on the CityPersons dataset. These results indicate that the model excelled in detecting pedestrians in urban environments, with a slightly higher detection accuracy on the Caltech benchmark compared to the CityPersons dataset.

In recent studies on pedestrian detection, Liu et al. has introduced a technique known as Center and Scale Prediction (CSP). Liu et al manage to advance the current state-of-the-art in pedestrian detection by proposing a new perspective in high-level semantic feature detection CSP. The authors undertake experiments on two of the most popular pedestrian datasets, the Caltech and CityPersons datasets. The Caltech dataset contains around 2.5 hours of footage taken from autonomous driving, which is extensively labelled. The CityPersons dataset is a subset of the Cityscapes dataset and contains 2975 training images, 500 validation images and 1575 testing images, with an average of 7 pedestrians in each. The authors chose these datasets since they provide bounding boxes which align well with the centres of pedestrians. The proposed solution comprises of two components, the feature extraction, and the detection head. The feature extraction modules concatenates feature maps of different resolutions into a single map which is then fed to a 3x3 convolutional layer in the detection head, followed by two prediction layers. The prediction layers focus on the centre location and the corresponding scale. The output is a centre and heat map. The method was implemented in Keras using ResNet and ImageNet as the backbone architectures. The adam network optimizer was selected. Training for CalTech was done in mini-batch of 16 images on a GTX 1080Ti GPU with a learning rate of 0.001, stopping at 15,000 iterations. Training for CityPersons was done using a learning rate of 0.0002 with 2 images per GPU, utilising 4 GPUs and stopping at 37,500 iterations. The proposed system advanced the current state-of-the-art by reducing the miss rate on the Caltech dataset to 4.5% from 5.0% achieved by RepLoss. Experiments on CityPersons reduced the miss rate to 3.8% from 4% of RepLoss and 4.1% of OR-CNN. The authors have published their work with code on a Git repository which has also been integrated into the Pedestron (<https://github.com/hasanirtiza/Pedestron)> project.

Zhang et al. conducted a study on pedestrian detection using deep learning technique, while also bringing up critical role’s computer vision can take up in today’s world to enhance safety. However, it introduces new deep learning techniques with the approach of CSANet in the aim of incorporating dual attention mechanisms to enhance the representation of feature maps. Another particular aim of this paper is to improve on AdaptFasterR-CNN which uses the CityPersons dataset to train the model, which was designed to utilize the dataset for strong generation capabilities. CSANet’s proposed method is targeted to be anchorfree, being unrestricted by a predefined anchor box ratio and instead attempts to predict bounding boxes and key points to make up objects. A study was conducted in aim to evaluate CSANet using the CityPersons dataset, analysing its components like feature extraction, fusion, and attention modules. Results show that in the experiment a NVIDIA GTX 1080Ti was used with 2 mini batchs of images were used from the CityPersons dataset which resulted in a unique test time of 270ms, meaning CSANET came second to the CSP model that was trained with batch 8. Integration of CAM (Channel Attention Module) and SAM (Spatial Attention Module) into the ResNet-50 backbone of CSANet which serves as an enhancement feature in the aim of improving performance of the model by effectively enhancing high-level semantic features and the ability to capture long range dependencies with feature maps. With the integration of CAM and SAM into the backbone of CSANet, the model can effectively capture both channel-wise and spatial attention, leading to improved pedestrian detection performance.

A common challenge researchers face in detection is accurately maintaining bounding boxes around the intended target. Liu et al. sought to address the challenge of pedestrian detection in crowded scenes by introducing the Adaptive-NMS algorithm. This algorithm refines bounding boxes based on target density. An experiment was conducted to validate the effectiveness of the Adaptive-NMS algorithm in enhancing pedestrian detection performance in crowded scenes using different datasets which can affect its deep learning algorithms to detect target pedestrians in a big crowd of people. They had used 2 datasets for their evaluations, CityPersons dataset and the CrowdHuman dataset. The CityPersons dataset had contained 5,000 images with annotations for approximately 35,000 labelled persons and 13,000 ignored region annotations while the CrowdHuman Dataset contained 15, 000,4,370 and 5,000 images that are found trough out the internet. The CrowdHuman dataset contained images with much denser populations compared to those in CityPersons. However, CityPersons encompassed a broader range of weather conditions across 18 cities in Germany with an average of 7 pedestrians in average per image. With a base learning rate set to and 0.002 for FPN and RFB and with 4 Titan X GPUs the results had shown that using this method with CrowdHuman dataset should return a higher detection rate for targets in large crowds.

So far, numerous datasets that have been created to tackle pedestrian detection across diverse weather conditions, varying light intensities, and complex environmental contexts. A standout dataset, the EuroCity Persons dataset, distinguishes itself in its specialization for traffic scenarios. Braun et al. had attempted to design the EuroCity Persons dataset to stand out as an invaluable resource for training models tailored to the complexities of detecting pedestrians amidst vehicular traffic. The EuroCity Persons dataset comprises over 238,200 manually labelled person instances in over 47,300 images, collected from vehicles across 31 European cities. Using this as a benchmark to train models to detect pedestrians they had experimented on Faster R-CNN by integrating the region proposal network (RPN) into Faster R-CNN, with aim of increased efficiency. They had utilized region proposal network (RPN) to generate region proposals directly from the convolutional feature maps. Experiments conducted with EuroCity Persons dataset demonstrated the effectiveness of the optimized Faster R-CNN in traffic scenarios and complex environments. To conduct such experiments an Intel(R) Core(TM) i7-5960X CPU with a GTX TITANX 12GB was utilized with the possibility of Improving the runtime by replacing the convolutional neural networks (CNN) VGG base architecture with the GoogLeNet model. The results had shown that the Faster R-CNN model achieved a log-average-miss-rate of 7.9, 17.0, and 33.2 on the "reasonable," "small," and "occluded" test cases, respectively. Despite having more training data in the EuroCity Persons dataset compared to other datasets with the deep learning methods not being able to fully utilizes the potential in detection leaving room for improvement.

**Literature Papers**

**High-level Semantic Feature Detection: ANewPerspective for Pedestrian Detection**

**CSANet: Channel and Spatial Mixed Attention CNN for Pedestrian Detection**

**CityPersons: A Diverse Dataset for Pedestrian Detection**

**Adaptive NMS: Refining Pedestrian Detection in a Crowd**

**EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes**