literature review

**High-level Semantic Feature Detection: A New Perspective for Pedestrian Detection**

This paper discusses a methodology for pedestrian detection based on high level semantic detection. Object detection often relies on a backbone of residual networks of deep learning models which take an image of the desired data as an input, reverting it into a map of different resolution using the formula φi = fi(φi−1) = fi(fi−1(...f2(f1(I)))) φi representing the maps output. Using a pretrained standard network on ImageNet (e.g. ResNet-50 and MobileNet) multi-scale feature maps can be created with each per stage/layers of convolutional neural network (CNN) and combined into one that contains more sematic information. An experiment was conducted using the Cross Stage Partial Network (CSP) focused on evaluating the performance of its two main components: the feature extraction module and the detection head. The featured extraction module, based on ResNet-50, was responsible for processing the input images and extracting feature maps with multiple resolutions. These feature maps were able to capture information from different scales, however a downside of this method (CSP) occurs in the simple design of the detection head. The detection head made of a single 3x3 layer followed by the prediction layers offers efficiency at the cost of limited ability to capture intricate object details and context in complicated scenarios. The experiment utilized the Caltech and CityPersons datasets for pedestrian detection. Caltech comprises approximately 2.5 hours of driving video footage with extensively labelled bounding boxes, while CityPersons is a challenging large-scale pedestrian detection dataset. The feature extraction module processed the data, and 4024 test frames were extracted for evaluation. The training process employed a ResNet-50 backbone pretrained on the ImageNet dataset, a common practice in deep learning. Optimization was performed using the Adam algorithm, commonly used for training neural networks and Mini-batch training was used for the divided dataset for a more efficient result.

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

This study addresses the same task of 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 Caltech dataset, analysing its components like feature extraction, fusion, and attention modules. 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.