# Pedestrian Detection Systems Focusing on Occluded and Small-Scale Individuals

Ameen Abdelmutalab

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy at

Concordia University

Montréal, Québec, Canada

December 2023

**CONCORDIA UNIVERSITY**
**SCHOOL OF GRADUATE STUDIES**

This is to certify that the thesis prepared

By: Ameen Abdelmutalab

Entitled: Pedestrian Detection Systems Focusing on Occluded and Small-Scale Individuals

and submitted in partial fulfillment of the requirements for the degree of

**Doctor Of Philosophy** (Electrical and Computer Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Rabin Raut — Chair

Dr. Jane You — External Examiner

Dr. Rajamohan Ganesan — Arm's Length Examiner

Dr. M. Omair Ahmad — Examiner

Dr. Wei-Ping Zhu — Examiner

Dr. Chunyan Wang — Thesis Supervisor (s)

Approved by

Dr. Yousef R. Shayan — Chair of Department or Graduate Program Director

December 14, 2023.
Date of Defence

Dr. Mourad Debbabi — Dean,

# Abstract


## Pedestrian Detection Systems Focusing on Occluded and Small-Scale Individuals

**Ameen Abdelmutalab, Ph.D.**

**Concordia University, 2023**

Pedestrian detection is essential in various applications, such as self-driving vehicles, video surveillance, and intelligent traffic management. However, the wide variations in pedestrian sizes, postures, locations, and backgrounds make the detection a complex task. In particular, the detection becomes significantly challenging due to the lack of pedestrian information when pedestrians are occluded by other objects, such as vehicles or trees, or when they appear as objects of small-scale in an input image. Such situations occur frequently in the real world. The objective of this thesis is to design CNN-based pedestrian detection models to improve the detection of occluded and small-scale pedestrians.

The first part of this work addresses the occlusion problem by proposing a specific detection model referred to as Multi-Branch Center and Scale Prediction (MB-CSP). The proposed model employs a multi-branch structure to optimize the utilization of the features extracted from the visible parts of pedestrians. This structure enables the feature data from the upper, middle, and lower parts of a pedestrian, as well as those of the full body, to be processed separately. By doing so, the data representing the true pedestrian appearances, whether partially or fully visible, can be more dominating in the final decision making. As a result, the interference from non-pedestrian data in the detection can be minimized. To optimize the fusion of the detection outcomes

generated by the multiple branches, a new method referred to as Boosted Identity Aware-Non Maximum Suppression (BIA-NMS) is developed and applied in the design of the MB-CSP detection system. The BIA-NMS method eliminates redundant detections across branches and boosts the scores of the preserved detections. To implement the proposed model, a part annotation algorithm has been introduced to enable the training of the multi-branch structure. It is anticipated that the proposed model can boost the overall performance of the pedestrian detection system.

The second part of this work provides a number of approaches to improving the detection of small-scale pedestrians, besides the occluded ones. One can use two CNNs designated to detect pedestrians of large and small scales, respectively, to achieve a good detection in each of the two cases. Instead of two designated CNNs, one can use only one and incorporate a specific branch in the proposed MB-CSP model to process the features of small-scale pedestrians. The other approach proposed in this thesis is to segment the original input image into multiple partially overlapped sub-images, the likelihood of the presence of small-scale pedestrians in each sub-image is measured, and those of high scores are selected and enlarged. The detection is performed by two CNNs, of which one is designed for the original image and the other for the selected/enlarged sub-images, in order to enhance the detection of small-scale pedestrians while preserving the detection quality of the occluded pedestrians.

The detection systems presented in this thesis have been trained and evaluated using image samples from the Caltech-USA and CityPersons datasets. The tests have confirmed the effectiveness of the proposed multi-branch system in detecting occluded pedestrians. The test results have also demonstrated that the approaches to enhance the small-scale pedestrian detection produced a visible improvement in this aspect without affecting the detection of occluded pedestrians.

# Contents

# List of Figures

# List of Tables

# List of Symbols

| | |
|---|---|
| $\boldsymbol{\alpha_i}$ | Weight of the loss corresponding to branch $i$. |
| $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ | Hyper-parameters of the loss of the center map. |
| $\boldsymbol{BB_i}$ | Bounding box for pedestrian part $i$. |
| $\mathbf{B}_{\boldsymbol{duplicate}}$ | The list of duplicated boxes. |
| $\mathbf{B}_{\boldsymbol{max}}$ | The detected box with the highest score. |
| **H, W** | Input Image height and width, respectively. |
| $\mathbf{h}_{\boldsymbol{n}}$, $\mathbf{w}_{\boldsymbol{n}}$ | Height and width of pedestrian n. |
| $\mathbf{H}_{\boldsymbol{Enlarged}}$ | The height of the regions selected for enlargement. |
| **IoU(A, B)** | Intersection Over Union for bounding boxes A and B. |
| $\boldsymbol{\lambda}$ | The boosting weight of the BIA-NMS method. |
| $\mathbf{Loss}_{C\_i}$, $\mathbf{Loss}_{S\_i}$, $\mathbf{Loss}_{O\_i}$ | Center, Scale, and Offset losses of branch $i$. |
| $\mathbf{Loss}_i$ | Total Loss of branch $i$. |
| $\mathbf{Loss}_{\boldsymbol{T}}$ | Total Loss of all the Branches. |
| $\mathbf{M}_{\boldsymbol{ij}}$ | A 2d Gaussian map presented at location $i, j$. |
| $\mathbf{MR}^{\mathbf{-2}}$ | The log-average miss rate. |
| **N** | The number of duplicated boxes detected by different branches. |
| $\mathbf{N}_{\ \boldsymbol{obj}}$ | The number of objects (specific body part) in the image. |
| $\boldsymbol{p_{ij}}$ | The predicted probability for a center to be present at location $(i, j)$. |

| | |
|---|---|
| $\mathbf{r}$ | Input image downsizing rate. |
| $\mathbf{Ratio_1}$ | The ratio between the standard deviation of the height distribution for pedestrians over 40 pixels and input image height. |
| $\mathbf{row}_{mean1}$ | The central row in the distribution of pedestrians over 40 pixels height. |
| $\mathbf{row}_{mean2}$ | The central row in the distribution of pedestrians between 30 and 80 pixels height. |
| $\mathbf{s}_n$, $\mathbf{t}_n$ | The network's prediction and the ground truth for each positive sample, respectively. |
| $\mathbf{Score}_B$ | Boosted detection box score. |
| $\mathbf{Score}_O$ | Old detection box score. |
| $\mathbf{std_1}$ | The standard deviation of the height distribution for pedestrians over 40 pixels. |
| $\mathbf{std_2}$ | The standard deviation of the height distribution for pedestrians between 30 and 80 pixels height. |
| $\boldsymbol{w_f}$, $\boldsymbol{h_f}$ | Width and height of the full pedestrian box. |
| $\boldsymbol{w_v}$, $\boldsymbol{h_v}$ | Width and height of the visible pedestrian box. |
| $\mathbf{x}_n$, $\mathbf{y}_n$ | Coordinates of the center of pedestrian n. |
| $\boldsymbol{x_f}$, $\boldsymbol{y_f}$ | Coordinates of the top left corner of the full box. |
| $\boldsymbol{x_v}$, $\boldsymbol{y_v}$ | Coordinates of the top left corner of the visible box. |

# List of Abbreviations

| | |
|---|---|
| **BIA-NMS** | Boosted Identity Aware Non-Maximum Suppression |
| **CNNs** | Convolutional Neural Networks |
| **CSP** | Centre and Scale Prediction |
| **DR-CNN** | Deconvolution Region-Based CNN |
| **FLOPs** | Floating-Point Operations Per second |
| **FPN** | Feature Pyramid Network |
| **HOG** | Histogram of Oriented Gradients |
| **ICF** | Integral Channel Features |
| **IoU** | Intersection over Union |
| **MB-CSP** | Multi-Branch Center and Scale Prediction |
| **MDSSD** | Multi-Scale Deconvolutional Single Shot Detector |
| **NMS** | Non-Maximum Suppression |
| **R-CNN** | Region-Based CNN |
| **RoI** | Region of Interest |
| **RPN** | Regions Proposal Network |
| **RUN-HM** | RegionUpscaleNet-HeatMap |
| **SAF R-CNN** | Scale-Aware Fast R-CNN |
| **SIFT** | Scale Invariant Feature Transform |
| **SOAM** | Scale-Specific Objectness Attention Mechanism |
| **SSD** | Single Shot Detector |

| | |
|---|---|
| **SVM** | Support Vector Machines |
| **VJ** | Viola-Jones |
| **YOLO** | You Only Look Once |

# Chapter 1

# Introduction

## 1.1 General

Pedestrian detection is an important part of various automatic surveillance systems. The reliable detection performance is critical for such a system to be used in practice. It is, however, very challenging to achieve because of the variability of pedestrians' appearances.

- Pedestrians can appear in many different ways and their postures can vary greatly. They may be walking, standing, or even cycling. This variability in appearance and posture makes it challenging to develop a pedestrian detection model that can accurately identify pedestrians in various situations.

- Various illumination conditions and environmental factors, such as day and night lighting, rain, storms, and snow, can significantly alter a pedestrian's appearance. These changes complicate the detection process by affecting visibility and the detection system accuracy.

Given the general challenges associated with the pedestrian detection, the work in this thesis specifically addresses two critical challenges. Firstly, the pedestrian

Figure 1.1: Image from CityPersons dataset [1], showing great variations in pedestrian's appearances including clothing, illuminations, scales, and postures. It also presents pedestrians with different occlusion patterns such as heavy occlusion, partial occlusion, and non-occlusion, as well as inter-class occlusion and intra-class occlusion.

detection becomes even more challenging when part of a pedestrian appearance is hidden by an external object, such as a tree or parked vehicle, which is referred to as inter-class occlusion, or by another individual, referred to as intra-class occlusion. In such scenarios, the visual features of the pedestrian are mixed with those of the obstructing objects, making the detection more complicated. This issue should be properly addressed, and effective solutions are needed. The second challenge is to detect small-scale pedestrians. Due to their sizes in images, their visual features are very hard to extract and to identify. Fig. 1.1 depicts actual street scene to illustrate various pedestrians' appearances, including occluded and small-scale pedestrians.

Recent technological advancements have significantly facilitated the pedestrian detection systems. This is driven by two key factors. Firstly, the growth in computational power allows for complex calculations that are essential for accurate pedestrian detection. Secondly, the increase in data storage capacity and the availability of

large amount of data, facilitates more comprehensive analyses, thereby boosting the efficiency of these systems.

## 1.2 Literature Review

Pedestrian detection methods can generally be classified into two main groups based on the approach used for detection. The first group involves the use of engineered features in combination with classical classifiers. In this approach, pedestrian features are manually extracted from images or videos, and then fed into a machine learning classifier to distinguish between pedestrians and non-pedestrians. The second group involves the use of deep learning models with automatic features. This approach utilizes deep neural networks to automatically learn and extract features from the images or videos, and then classifies the objects in the scene. While both approaches have their advantages and limitations, the development of deep learning models has revolutionized pedestrian detection, allowing for greater accuracy and reliability in detecting pedestrians in various scenarios.

### 1.2.1 Pedestrian Detection Using Engineered Features

Engineered features are defined by researchers based on their observations and analysis of a particular problem. These features are extracted to better distinguish between different objects, in this case, pedestrians from other objects in images. In [2], one of the earliest proposed methods for pedestrian detection utilized Haar-like features. Haar-like features are simple rectangular filters that can be used to compute local image contrast. These filters are applied to an image to identify distinctive features such as edges by computing the difference between the sum of pixel intensities in white and black rectangles of the same size and shape. The Viola-Jones (VJ) algorithm, described in [3], accelerated the computation of Haar-like features by utilizing the

Integral Image approach. This approach involves pre-calculating the sum of pixel intensities in rectangular regions of an image, which allows for efficient computation of Haar-like features. The approach also combined Haar-like features with Cascaded Ada-boost classifiers, resulting in improved detection accuracy. Besides Haar-like features, Scale Invariant Feature Transform (SIFT) is another widely used feature extraction technique for pedestrian detection. SIFT features are known for their robustness to rotation and scale changes, enabling them to detect pedestrians in images captured under various viewing conditions. The SIFT method identifies and characterizes key points in an image using a set of descriptors that are invariant to changes in scale. These descriptors can be compared across multiple images, facilitating reliable pedestrian detection, and tracking in complex scenarios. Furthermore, the Histogram of Oriented Gradients (HOG) method was developed by Dalal and Triggs in their paper [4]. This method entails partitioning an image into various blocks and computing the gradient histogram based on the magnitudes and angles of the image gradients. The resultant histogram depicts the spatial distribution of edge orientations within the image and can serve as a useful feature vector for pedestrian detection. Additionally, the authors of [5] presented the Integral Channel Features (ICF) algorithm as an accurate and computationally efficient technique for pedestrian detection. The ICF algorithm generates a set of feature channels using integral images, which approximate detection features such as HOG, color statistics, and linear filters at different scales by estimating their values from neighbouring scales. The resulting feature channels capture both local and global information about pedestrians, which are then processed by a cascade of classifiers to detect pedestrians in the image.

Classical machine learning classifiers, including Support Vector Machines (SVM) [4, 6, 7] and boosting techniques [3, 8–10], are commonly used in the literature with engineered features to achieve better detection performance. SVM is a popular classifier that separates data into different classes using a hyperplane where this hyperplane is

chosen such that it maximizes the margin between the closest points of the different classes. Boosting techniques, on the other hand, are used to improve the performance of weak classifiers by combining them into a strong classifier. Boosting classifiers work by sequentially training a series of weak classifiers on the same dataset, with each subsequent classifier focusing more on the misclassified examples of the previous classifiers. The final classifier is obtained by combining the outputs of all the weak classifiers, weighted according to their individual performances.

Recent research in the field of pedestrian detection has focused on the evaluation and comparison of different state-of-the-art detectors. For example, Dollar et al. conducted a systematic analysis in [11] focused on 16 state-of-the-art detectors across six datasets, evaluating and comparing the performance of these detectors in a unified manner at different scales and occlusion patterns. In another study, Benenson et al. analyzed over 40 pedestrian detectors on the Caltech-USA dataset [12]. Additionally, a survey in [13] reviewed 30 methods from the past decade, highlighting recent advancements in pedestrian detection. Such studies helped researchers to improve detection accuracy and reliability.

### 1.2.2 Pedestrian Detection Using CNN Features

Convolutional Neural Networks (CNNs) are now widely used in pedestrian detection systems. There are two main approaches to using CNN models, hybrid approach and pure CNN-based approach. Hybrid approach may use deep learning to extract features and then use traditional classifiers for decision making. An example of this, is the work proposed by Yang et al. [14], who replaced engineered features with those derived from convolutional layers. Alternatively, hybrid approach may merge conventional features with a CNN for the classification task. An instance of this latter approach is presented by Ribeiro et al. [15], who trained several deep networks with varying inputs, such as colour and segmentation images.

Next method fully depends on deep learning for extracting features and performing classification. For example, Zhu et al. [16] carefully selected features from the Region of Interest (RoI) for regression and combined these features from various layers for classifying objects. These techniques, which are solely based on CNNs, have been found to be more efficient and simpler than hybrid methods, typically using an end-to-end training approach. Furthermore, Lin et al. [17] adopted a top-down structure to merge features from both deep and shallow layers, an approach referred to as Feature Pyramid Network (FPN). This demonstrates the ease and effectiveness of using a purely CNN-based approach in pedestrian detection.

Networks like Region-Based CNN (R-CNN) [18] employ a region proposal technique known as Selective Search method that uses different image features like brightness, color, texture, composition, and hierarchical structure to suggest proposed regions, these regions are then classified and adjusted using a CNN network. Fast R-CNN [19], on the other hand, directly applies the input image to a CNN network and generate the RoI from the feature maps. This approach considerably reduces the processing time compared to the original R-CNN, since the CNN network is only applied once. To further reduce training and testing time, Faster R-CNN [20] replaces classical methods for calculating region proposals like Selective Search with a Regions Proposal Network (RPN). This approach eliminates the need for external region proposal methods, and significantly reduces the computational complexity of the process compared to Fast R-CNN. The RPN generates region proposals in parallel with the network's classification and bounding box regression tasks, leading to faster and more accurate pedestrian detection performance.

An alternative approach to pedestrian detection utilizes feed-forward networks that do not require a specific network for region proposals. Methods in this category include Single Shot Detector (SSD) [21], You Only Look Once (YOLO) [22], and Centre and Scale Prediction (CSP) [23]. SSD generates anchor boxes of different sizes

and aspect ratios centred at every pixel in the input image. It then classifies these anchor boxes into pedestrian or backgrounds based on their aspect ratios. YOLO, on the other hand, generates non-overlapping anchor boxes and each box predicts a certain number of bounding boxes. These bounding boxes are then used to represent the position and size of pedestrians in the image. Both SSD and YOLO demonstrate high detection accuracy and fast processing times, making them suitable for real-time applications such as autonomous driving and surveillance. CSP is another feed-forward neural network that does not rely on a region proposals network to detect pedestrians. Instead, it automatically generates heat-maps that indicate the predicted locations of pedestrians. CSP has demonstrated state-of-the-art accuracy while maintaining high processing detection speeds.

### 1.2.3 Existing Solutions for Detecting Occluded Pedestrians

Occlusion is a common issue in pedestrian detection; in [11], videos recorded from a driving car that captured pedestrians in different cities, showed that 70% of pedestrians were occluded in at least one time frame. While this study was conducted in the greater Los Angeles region of the United States, it provides an example of the scope of the problem in similar metropolitan areas.

Occlusion occurs when part of pedestrian body is invisible and covered by another object, in general occlusion can be divided into two types, inter-class occlusion that takes place when pedestrians body is covered by an obstacle such as a tree, car or a suitcase, and intra-class occlusion which occurs when part of pedestrian is covered by another pedestrian in the scene, usually in crowded areas. Based on the degree of occlusion, pedestrians can be categorized into heavy occluded pedestrians, partially occluded pedestrians and non-occluded pedestrians. The great variation in occlusion patterns, as well as the occluding objects, makes it difficult for machine learning algorithms to learn a general model. Approaches to handle occlusion can be divided

into three categories, Part-Based approaches, Attention-Based approaches and Post-Processing and Loss-Based approaches. For models that utilize a Parts-Based approach, as highlighted in references [24–26], the full pedestrian body is divided into various parts, often characterized by different occlusion patterns. During occlusion some body parts remain visible, hence detecting these parts is more convenient compared to detecting the full body with mixed features of pedestrian and barrier. Earlier Part-Based approaches used ensembles models, in which separate parts detectors are used independently. As mentioned in [27], this approach is unsuitable for real-time processing where the system complexity grows linearly with the addition of every part detector. Moreover, ensembles models ignore the correlation between different parts during learning, resulting in a non context-aware parts detector. Other methods built parts models using joint frame work [28], where different body parts are trained collaboratively using single Convolutional Neural Network (CNN). This approach reduces the complexity presented in ensemble models; however it lacks accurate parts annotation. In [27], authors introduced Multi Label Learning with separate labels assigned to different body parts, their approach uses part pool with 20 different parts and classical Ada-boost classifiers. Authors in [29] introduced Visible-to-Full body Network (V2F-Net), in which the visible pedestrians are first identified and then used to estimate their full body extension. In the study by Noh et al. [30], the researchers used the confidence of pedestrian's visible parts to adjust the final detection confidence. This method addressed the issue of low confidence when pedestrians are partially occluded.

As for Attention-Based approaches, authors in [31], introduced a separate part-attention network to Faster R-CNN with the objective of creating a channel-wise attention vector. This vector is used to adjust the weights of channel features to handle different occlusion patterns. Another example is introduced by Guo et al. [32], in their study to use a semantic segmentation map. These maps, derived from depth

images, guide the adjustment of the convolutional features obtained from RGB images. Furthermore, Lin et al. [33] utilize pedestrian attention masks that are aware of scale differences and a zoom-in-zoom-out module to enhance the feature maps' ability to detect smaller and partially hidden pedestrians.

For Post-Processing and Loss-Based approaches, the studies in [34] and [35] investigate the impact of the Non-Maximum Suppression (NMS) threshold on crowded detection. To mitigate the influence of a rigid threshold on detection, advanced NMS strategies are proposed in [36–39]. Soft NMS [40] aims to reduce the score of closely overlapping proposals rather than removing them, but it still indiscriminately penalizes boxes with high overlap. Certain studies incorporate additional information such as density and diversity into NMS, to address inflexible thresholds. Adaptive NMS [36] employs the larger value between the predicted density around the instance and the initial threshold as the dynamic suppression threshold to improve the bounding boxes. This implies that the threshold increases as instances overlap and decreases when instances appear independently. Different track of work focuses on improving crowded pedestrian detection by introducing new loss functions [34, 41], their goal is to minimize the distance between duplicate detection boxes of the same pedestrian and to maximize the distance between adjacent pedestrian boxes, eventually preventing over elimination by NMS. Other authors integrated additional innovative features to improve pedestrian detection task, for example Du et al [42] applied features from a pixel-wise semantic segmentation network, and Song et al [43] integrated temporal information from adjacent frames.

### 1.2.4 Existing Solutions for Detecting Small-Scale Pedestrians

Detecting larger pedestrians is generally easier for many models, but identifying smaller ones can be very challenging. This is generally due to several factors. (a) Small-scale pedestrians are described by a limited number of pixels in the image,

making it particularly challenging for models to extract significant features to identify them. (b) The blurring effect is severe on small-scale pedestrians, resulting in difficulty distinguishing the target from the background. (c) Pre-trained deep learning models with consecutive pooling layers, are optimized to detect objects that are usually large or moderate in size, the mismatch between these scales and small pedestrian's scales limits their performance.

Small-scale pedestrian detection methods are mainly classified into four pillars, Scale-Specific Categorization, Contextual Information, Super-Resolution, and Region-Proposal, as mentioned in [44]. Scale-Specific Categorization combines the pedestrian information extracted from shallow convolutional layers, which is usually important for small-scale pedestrian localization, with deep convolutional layers necessary for semantic information, leading to better pedestrian detection. For example, Multi-Scale Deconvolutional Single Shot Detector (MDSSD), proposed in [45], uses skip connection to add more contextual features. In this model, deconvolution layers are applied to upsample the high-level feature maps to the same resolution as the corresponding low layers. Deconvolution Region-Based CNN (DR-CNN) [46], Unlike MDSSD, which sums the deconvolution layers, DR-CNN concatenates them. It also introduces a new loss function to facilitate the training of hard negative samples, improving the model's overall performance. Authors in [47] developed Scale-Aware Fast R-CNN (SAF R-CNN) with two detection ends to recognize large-scale and small-scale pedestrians separately. Their model uses the later VGG layers to detect small-scale pedestrians, which is insufficient as these layers are optimized to detect large areas in the image.

Contextual Information, this approach makes use of the surroundings and environment around pedestrians to enhance the detection accuracy of small-scale pedestrians. For large-scale pedestrians, the features extracted are usually enough for recognition, but with small-scale pedestrians extracting additional supplementary information to complement the original features becomes essential. ContextNet [48] introduces a

novel region proposal network (RPN) designed to encode the context information surrounding a small-scale object proposal. Inside-Outside Net (ION) [49] utilizes spatial recurrent neural networks (RNNs) to search for contextual information outside the target region. This model integrates multiple scales and context information, enhancing detection capabilities.

Super-Resolution techniques work to convert raw low-resolution images into higher resolution versions. This means that more details of small-scale pedestrians can be obtained, improving the clarity and understanding of the images. For Perceptual GAN [50], a new conditional generator was introduced, utilizing low-level features as input to capture more details about small-scale objects, leading to a super-resolved representation. SOD-MTGAN [51] is a novel multitask generative adversarial network that produces super-resolved images with real high-resolution, containing high-frequency details. This results in easier classification and improved localization. JCS-Net [52] comprises two subnetworks, a classification sub-network and a super-resolution sub-network. These are integrated into a unified network by combining both classification loss and super-resolution loss, enhancing small-scale pedestrian detection.

Region-Proposal is a strategy aimed at creating suitable anchors for small-scale pedestrians. By focusing on specific needs rather than generic anchor parameters, it seeks to better fit small-scale pedestrians and improve detection accuracy. For example, AttentionMask [53] was designed to create tailored region proposals for small-scale objects. It adopts a Scale-Specific Objectness Attention Mechanism (SOAM) to select the most promising windows at each feature map with different scales, thereby reducing the number of sampled windows. While all scales are evaluated according to their attention values to find optimal locations for window sampling, this strategy focuses only on the most promising windows. As a result, it saves memory and GPU resources, enhancing the detection of small-scale objects. In addition, the authors in [54] employ oversampling techniques specifically for images with small-scale targets.

This approach is used to enhance the model's ability to predict small-scale targets accurately.

## 1.3 Objectives and Organization of the Thesis

The objectives of this thesis are two folds. The first fold is to develop a pedestrian detection system capable of enhancing occluded pedestrian detection, addressing both intra-class and inter-class occlusion challenges, all while ensuring a high level of simplicity and speed. The second fold is to design detection architectures that enhance the detection of small-scale pedestrians, while preserving the improved accuracy achieved in detecting occluded and large-scale pedestrians.

To achieve these objectives, the work presented in this thesis explore the following avenues:

**Multi-Branch Model for Occluded Pedestrian Detection**. The goal is to optimize the detection of occluded pedestrians by using dedicated branches for distinct body parts. Each potential pedestrian body is divided and the feature data of each part are directed to a designated branch. It attempts to process efficiently the features of the exposed parts of a pedestrian object and to ensure that the patterns of these parts are correctly recognized. In this way, the data produced from the exposed parts in these branches can significantly influence the final decision-making. The body partitioning should be done appropriately to ensure that each part has easily distinguishable patterns and the number of parts is minimal to minimize the overall model complexity.

**Optimized Fusion of the Multi-branch Model Detections**. To better handle the data produced by the multiple branches of the detection model, a post-processing method must be integrated. The purpose of this method is to eliminate duplicate

detections, optimize detection scores, and improve the detection of highly occluded pedestrians.

**Architectures to Address Pedestrian Size Variability**. The objective is to detect large-scale and small-scale pedestrians independently, using either distinct detection models or separate branches within a single model. Implementing these architectures can enhance the detection accuracy for smaller pedestrians without compromising the accuracy for larger ones.

**Architectures for Enlarging Potential Pedestrian Regions**. The objective is to identify image regions potentially containing small-scale pedestrians. These regions should then be enlarged and processed using a separate detection model to identify the pedestrians within them.

This thesis is organized as follows. In Chapter 2, a background overview essential for understanding the proposed models is provided. In Chapter 3, the proposed Multi-Branch detection model, designed to enhance occluded pedestrian detection, is introduced. Chapter 4 presents the architectures proposed to enhance small-scale pedestrian detection. Chapter 5 details the experimental setup and evaluates the proposed models. Finally, Chapter 6 concludes the thesis by summarizing its primary findings.

# Chapter 2

# Background

This chapter provides an overview of the essential components of a Centre and Scale Prediction (CSP) model. Throughout this thesis, systems based on the CSP model are proposed to enhance the detection of occluded and small-scale pedestrians. This choice is motivated by the CSP model's simplicity and high detection accuracy compared to other models in the literature. The chapter also covers the pre-processing techniques essential for system robustness, the architecture of the CSP model, and the post-processing method used to remove duplicate detections and produce the final detection outputs.

## 2.1 Pre-Processing Techniques

In the process of training pedestrian detection models, the utilization of limited datasets often leads to models with limited generalization capabilities, especially in the context of larger network architectures which have a higher risk of over-fitting. Recognizing these limitations, the application of image augmentation techniques is essential. These techniques aim to strengthen the model's stability, ensure a broader representation in the dataset, and effectively address over-fitting concerns. Key techniques in this field include.

1. Horizontal Flipping: Mirroring images horizontally adds variation to the dataset by changing the position of pedestrians, yet the inherent subject and context of the image remain unchanged.

2. Random Cropping: Selecting specific parts of images to create more variations, offering a broader view of pedestrian features.

3. Padding: By padding and then cropping images, pedestrians appear smaller, simulating varied distances or perspectives, enhancing model detection capabilities in diverse scenarios.

4. Image Resizing: Standardizes image sizes for neural networks to enhance the computational efficiency.

5. Random Noise: By adding noise to the training images, the model becomes better at detecting pedestrians in noisy real-world settings, increasing its robustness against diverse conditions.

## 2.2 Centre and Scale Prediction (CSP)

Anchor-based detection models primarily use predefined anchor boxes with specific scales and aspect ratios to identify objects of key importance. These models employ classification methods to determine the object's precise class, and regression strategies to accurately locate objects and define their spatial dimensions. In contrast, the Centre and Scale Prediction (CSP) model [23], displayed in Fig. 2.1, introduces an anchor-free approach. This model identifies the object's class and its spatial location directly from distinct image features. By avoiding the use of predefined anchor boxes, the CSP model provides a potentially enhanced flexibility in object localization, a fundamental aspect of computer vision tasks.


Figure 2.1: The architecture of the CSP model [23]. Including its Feature Extraction part, and Detection Heads part.

The CSP model consists of two main parts. Feature Extraction part and Detection Heads part. The Feature Extraction part, commonly known as the model's backbone, processes the input image and extracts essential detection features using series of convolutional layers and pooling steps. In this thesis, the ResNet50 architecture, detailed in Table 2.1, is the adopted backbone for the proposed CSP-based systems. The key strength of the Resnet architecture lies in its innovative use of residual connections. These connections address major deep learning challenges like, vanishing gradients, feature reuse, training speed, and model interpretability.

Table 2.1: ResNet50 Architecture Details.

| **Layer Name** | **Output Size** | **50-Layers** |
|---|---|---|
| Conv1 | $112 \times 112$ | $7 \times 7$, 64 filters, Stride 2 |
| Max pool | $56 \times 56$ | $3 \times 3$, stride 2 |
| conv2_x | $56 \times 56$ | $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix}$ x3 |
| conv3_x | $28 \times 28$ | $\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix}$ x4 |
| conv4_x | $14 \times 14$ | $\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix}$ x6 |
| conv5_x | $7 \times 7$ | $\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix}$ x3 |
| Avg pool | $1 \times 1$ | 1000-d fc, softmax |

Following feature extraction, the Detection Head part processes these features to produce three different heat-maps, namely the center, scale, and offset heat-maps.

**The Center Heat-Map** is an essential component of the pedestrian detection process using the CSP model, its role is to locate a pedestrian's center within an image. In the training phase, the center heat-map pixels are set to one if they represent the

center of a pedestrian, and to zero if they do not. However, the training process is complicated due to the substantial imbalance between the number of center and non-center pixels. To address this challenge, specific adjustments are made to the center heat-map loss function calculations, with a particular focus on utilizing the focal loss, as detailed in [55]. Focal loss is a specialized function designed to handle class imbalances as in pedestrian detection scenarios. Its role is to enhance the model's attention to challenging cases (pedestrians' center pixels) while reducing the impact of straightforward examples (non pedestrian pixels). Within this strategy, a 2D Gaussian distribution is applied around each pedestrian's center. So, even though pedestrian centers are marked as ones and non-pedestrian areas as zeros, the system offers some flexibility for nearby pixels by not setting them exactly to zeros during training. This technique simplifies the center heat-map training process, enhances the detection of challenging positive pixels, and aids in efficient model convergence.

**The Scale Heat-Map** is the second output heat-map and its primary purpose is to determine pedestrians' height ($h$) and width ($w$). In practice, measuring only the height is adequate, as pedestrians typically exhibit a consistent shape or aspect ratio. For example, when assigning a ground truth value to a location in the scale heat-map, if this location is the center of pedestrian $n$, then pixel in the scale heat-map is assigned the value $\log(h_n)$, corresponding to the log of the height of this pedestrian. To reduce ambiguity, $\log(h_n)$ is also assigned to negative locations within a radius of 2 from the pedestrians centers, while all other locations are assigned values of zero.

**The Offset Heat-Map** is the third output heat-map and its main function is to correct potential inaccuracies that may arise when detection occurs at a lower resolution than the original image size, which can lead to deviations or misplacements in detection. For each pixel location, two offset heat-maps are created. One for the horizontal shift to the pedestrian center point and another for its vertical shift. During

training, these offset heat-maps have zero values everywhere except at pedestrian locations, where they are assigned values as follows.

$$x_n - \left\lfloor \frac{x_n}{r} \right\rfloor, \quad y_n - \left\lfloor \frac{y_n}{r} \right\rfloor \tag{2.1}$$

where $r$ is the downsampling factor equals to 4 in a standard CSP setting. $x_n$ and $y_n$ represent the coordinates of the center of pedestrian $n$.

## 2.3 Non-Maximum Suppression (NMS)

During the model evaluation phase, the CSP model utilizes center maps to identify potential pedestrian center points. This is achieved by applying a threshold to pixel values. When a pixel's value exceeds the specified threshold, it is considered as a pedestrian center. In practical scenarios, several pixels surrounding the actual pedestrian center often display values surpassing the predefined threshold. This situation occurs due to inherent similarities among these pixel characteristics, further influenced by the model's training process involving the Gaussian mask, which allows for errors. As a result, multiple pixels are designated as centers for the same pedestrian, resulting in the generation of multiple overlapping bounding boxes, all indicating the presence of the same pedestrian.

NMS is a technique employed to manage multiple bounding boxes, each associated with its respective confidence score. By utilizing the concept of Intersection over Union (IoU), NMS selectively eliminates redundant bounding boxes with an IoU exceeding a specified threshold, often set at 0.5, only the bounding box with the highest confidence score is retained, a process visually depicted in Fig. 2.2. The underlying assumption here is the absence of pedestrians in immediate proximity, a scenario that would result in bounding boxes with IoU values surpassing the predefined threshold. In this context, the IoU between two boxes A and B is mathematically defined as follows.

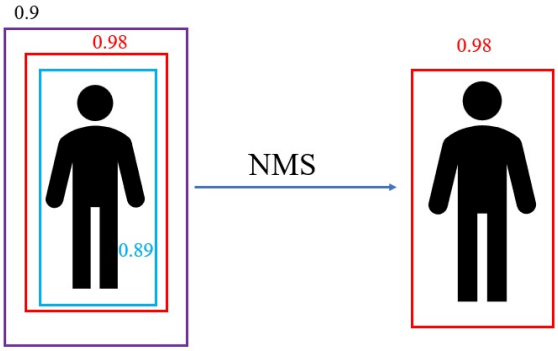



Figure 2.2: Illustration of Non-Maximum Suppression (NMS) in action, showing the removal of redundant bounding boxes based on the IoU criteria. The assumption here is that the blue and purple boxes have an IoU of more than 0.5 with the red box (the box with the highest confidence), resulting in their removal when applying the NMS method.

$$IoU(A, B) = \frac{\text{Area of Intersection}(A, B)}{\text{Area of Union}(A, B)} \tag{2.2}$$

## 2.4 Summary

In this chapter, the main components of the Centre and Scale Prediction (CSP) model were presented. Different image augmentation techniques are introduced to address the challenges associated with limited datasets and over-fitting in models. The anchor-free nature of the CSP model was compared to traditional anchor-based models, with the ResNet50 architecture serving as its backbone. The roles of the center, scale, and offset heat-maps in detection were described. The chapter also explored the Non-Maximum Suppression (NMS) technique and its use of the Intersection over Union (IoU) metric to eliminate redundant bounding boxes.

# Chapter 3

# Proposed CNN Model for Enhanced Occluded Pedestrian Detection

In this chapter, a Convolutional Neural Network (CNN) system, labelled as Multi-Branch Center and Scale Prediction (MB-CSP) is presented. This system is specifically developed to tackle the challenges in detecting occluded pedestrians. When pedestrians are obscured, the clarity of the available data can be significantly reduced, making the detection more challenging. The strategy introduced in this chapter aims to use the available data efficiently to improve the detection accuracy without significantly increasing the system's computational requirements. The work described in this chapter has been presented in the research paper entitled "Pedestrian Detection Using MB-CSP Model and Boosted Identity Aware Non-Maximum Suppression", published in IEEE Transactions on Intelligent Transportation Systems [56].

In Section 3.1, the challenges of detecting occluded pedestrians are introduced, and the building blocks of the proposed MB-CSP system are presented. Section 3.2 explains The Feature Generation Block, which is responsible for extracting the

significant features for pedestrian detection. The Upper-Middle-Lower and Full (UMLF) Block, designed to refine the generated features and detect pedestrian's body parts, is illustrated in Section 3.3. Section 3.4 introduces the post-processing Block designed to merge the multi-branch outputs. The algorithm developed for parts annotations is outlined in Section 3.5. Finally, Section 3.6 discusses the loss functions utilized in the proposed system.

## 3.1 Overview

Detecting individual pedestrians in crowded areas is a challenging task, as people are often occluded. A pedestrian can be partially obstructed by objects of other classes such as vehicles and trees, which is referred to as inter-class occlusion. An intra-class occlusion occurs when a pedestrian is partially occluded by other pedestrians. In general, there are two hurdles when detecting occluded pedestrians.

- Real pedestrian features are mixed with features of the occluding barrier. This hurdle is present in both inter-class and intra-class occlusions and can result in confusion when learning pedestrian characteristics, eventually leading to wrong detections. To overcome this hurdle, the proposed system utilizes part-based detectors, each of which is exclusively learned from visible pedestrian parts.

- Multiple detections of a single pedestrian is a common problem in most detection systems. The proposed Multi-Branch Center and Scale Prediction (MB-CSP) system may exacerbate this problem by creating duplicates from its different branches. To address this issue, the proposed system utilizes Non-Maximum Suppression (NMS) to eliminate duplicates within the same branch and proposes a novel post-processing method referred to as Boosted Identity Aware Non-Maximum Suppression (BIA-NMS) for removing duplications across the different branches.

The block diagram of the proposed MB-CSP system is illustrated in Fig. 3.1. It is composed of the following three blocks.

1. **Feature Generation Block:** This block processes input images to extract distinguishing features, producing feature maps optimized for pedestrian detection at multiple scales.

2. **UMLF Block:** The block divides feature data into four branches. Each branch creates heat-maps indicating location, scale, and offset of pedestrian targets. This four-branch method enhances detection in different scenarios, increasing reliability and capturing more pedestrians in varied conditions.

3. **Post-Processing Block:** This block integrates detections from the different UMLF branches, processing the combined information to remove duplications and to enhance the detection of intra-class occlusion cases.

Each block is discussed in detail in the following sections.

## 3.2 Feature Generation Block

The Feature Generation Block is fundamental to the pedestrian detection process. Its main task is to process the input images and extract the essential features for pedestrian detection. The block adopts ResNet50 architecture [57] and is pre-trained on the ImageNet dataset. This pre-training allows for more detailed feature extraction from a broad spectrum of image data.

The input images, sized HxW, are processed by the feature generation block. As they are passed through stages 3, 4, and 5, their sizes change to $(W/8)^2$, $(W/16)^2$, and $(W/16)^2$ respectively. To make them ready for the next steps, they are first refined by deconvolutional layers. Next, L2-normalization is applied to adjust their sizes to a more uniform format, specifically $(H/r) \times (W/r)$. It's important to mention that the

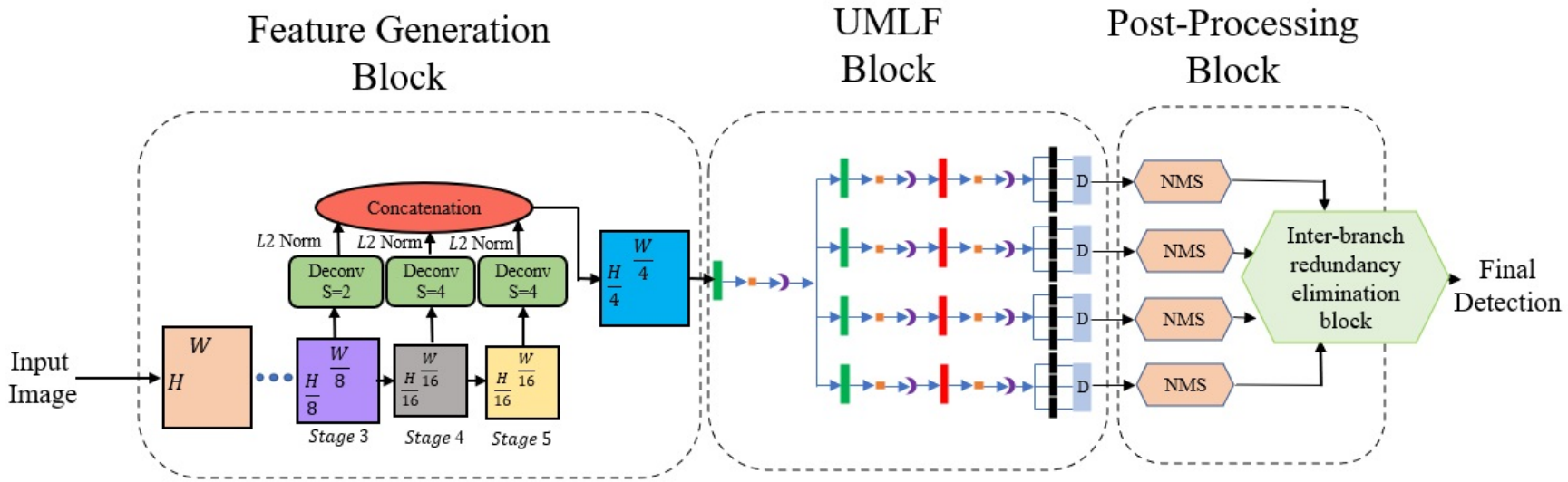



Figure 3.1: MB-CSP architecture consisting of three blocks: Feature Generation Block, UMLF Block, and Post-processing Block.

downsizing rate used is represented by $r$ and is set to 4. After these adjustments, the images are given detailed information from different levels. With these enhancements, they are then fed into the UMLF block for the main detection phase.

## 3.3 UMLF Block

If pedestrians appear partially occluded in an image, the pixel data in the occluded part carry the features of the occluding barriers, which can contribute adversely to the detection of the pedestrians. UMLF block is designed to mimic human perception of a partially occluded pedestrian by extracting the relevant features from visible pedestrian parts and ignoring data variations in the occluding barrier. To do so, one needs to partition the view of a pedestrian into parts so that visible areas containing actual features of real pedestrians are separated from the occluded areas, making it possible to exclude non-pedestrian features in the training process. As a result, the features of pedestrian parts will be processed in separate branches of the block, and each branch will learn features of its corresponding part.

A pedestrian appearance can be obstructed differently, and the patterns of occlusion are not unique. To partition a pedestrian view appropriately, the following elements are considered.

- The designed partitions must have recognizable and distinguishable patterns that discriminate pedestrians from irrelevant objects.
- Partitions must suit the different occlusion patterns so that each pedestrian has at least one visible part, without significant interference of occluding element, in most of the occlusion scenarios.
- The number of partitions must be reasonable, as more partitions implies more branches, therefore increasing the complexity of the overall system.


Figure 3.2: Pedestrians' parts in different occlusion scenarios: A) shows a fully visible pedestrian, where upper, middle, and lower body parts represent human contours. B) A partially occluded pedestrian, where the upper and middle parts carry pedestrian information, and the lower part is irrelevant. C) Depicts a heavily occluded pedestrian where only the upper part indicates the presence of a pedestrian.

As discussed earlier, a pedestrian's view has been divided into overlapping parts: upper, middle, lower, and full body parts. Fig. 3.2 depicts these parts under different occlusion conditions. Correspondingly, the MB-CSP system is constructed with four distinct branches, with each branch specifically designed to recognize and identify one of these parts, as follows:

- **Upper Part Branch.** This branch is dedicated to detecting pedestrians face and shoulders using their distinguishable contours. The upper part detection is crucial in detecting highly occluded pedestrians, where face and shoulders might be the only visible part.

- **Middle Part Branch.** The features of the middle part, including the torso, of a pedestrian's view are very different from the upper or lower parts. This branch is trained to identify the patterns of the middle part, and its output data help to detect reasonable and partially occluded pedestrians.

- **Lower Part Branch.** This branch is specialized to detect the unique shape of the lower part, i.e. the trunk and legs of a pedestrian. If this part is visible, this branch will detect it and contribute to the correct final decision.

- **Full Body Branch.** In case of fully visible pedestrians, a full-body detection is evidently more advantageous than that of part-based, particularly when there are many fully visible pedestrians in the training samples. Hence, this full-body branch is placed to minimize the risk of missing fully visible pedestrian targets.

A good pedestrian detection needs a good identification of the patterns distinguishing the pedestrian targets from the rest of the image. The four-branch structure of the proposed UMLF Block permits each branch to be trained specifically to identify the distinguished patterns of the designated part. If the part is visible, the branch will generate a significant output, otherwise, no target patterns will be detected and the output will be weaker. The final detection decision is based on the outputs of all the four branches, dominated by the data generated from the visible parts.

The detailed structure of the UMLF block is illustrated in Fig. 3.3. The input data, i.e., the 2D maps carrying features extracted in different scales, are first fused by means of a convolutional layer of 256 kernels. The outputs of this layer are then applied to each of the four branches for the detections of the upper, middle, lower and full-body parts, respectively.

In each of the four branches, as shown in Fig. 3.3, the detection of the designated part is performed by two convolutional layers, each of which has 256 kernels. It should be noted that a standard 3x3 convolution is applied in the first layer, whereas the second layer is a 3x3 separable convolution (consisting of depth-wise filter of size 3x3 followed by 1x1 classical convolution filter). The separable convolution acts as a channel-wise attention mechanism to highlight the important features in each map. The output data containing information on targets centers, scales and offsets are then processed by 1x1 convolutions to generate the final center, scale and offset heat-maps.

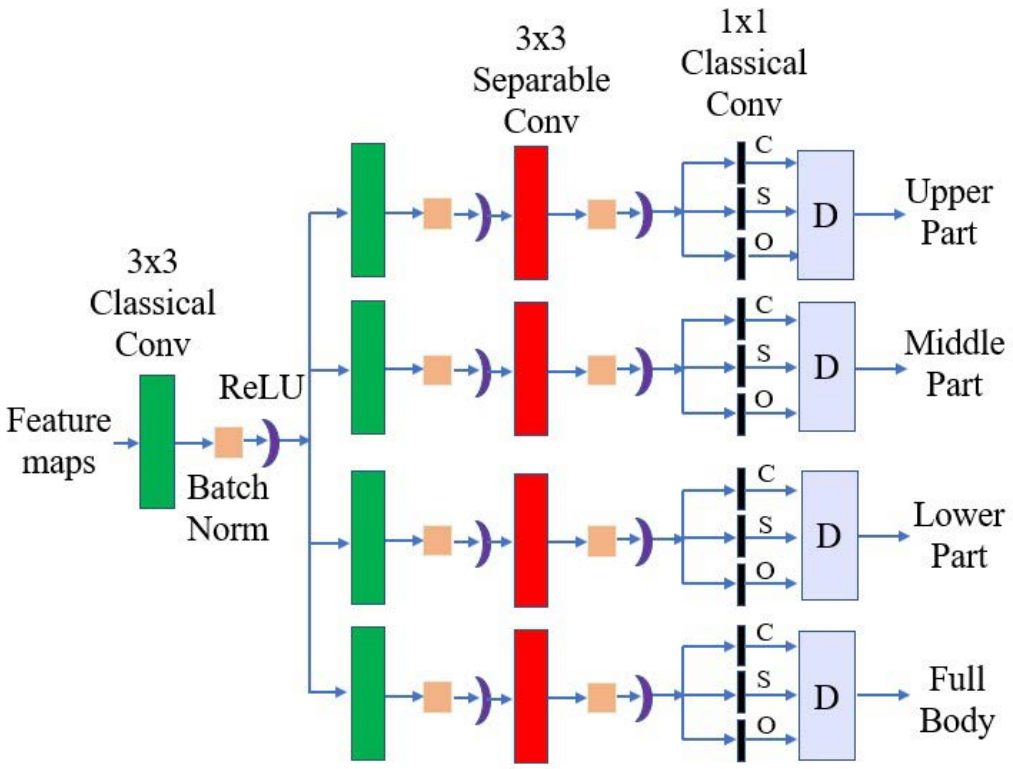
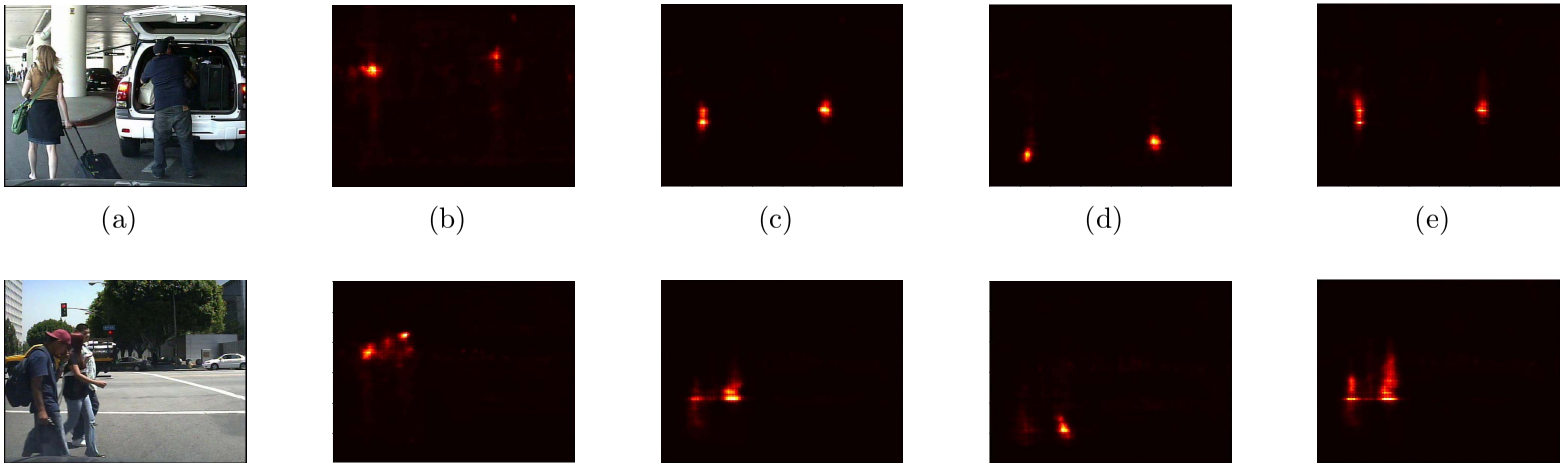



Figure 3.3: UMLF Network architecture, where C, S, O, and D denote center map, scale map, offset maps, and output decoder, respectively.

UMLF branches are configured identically. However, the convolution kernel parameters in each branch are designed to learn the associated features of each part. Fig. 3.4 illustrates two detection examples, each having an original input image and its associated upper, middle, lower and full-body center heat maps generated by the four UMLF branches. The first example involves two fully visible pedestrians, with their corresponding four center heat maps, indicating clearly and coherently the locations of their parts and full-bodies. The second example is a challenging heavy occlusion case, as one of the three pedestrians is severely occluded. Accordingly, the full-body branch can only detect two pedestrians, as shown in Fig. 3.4(j). So do the branches for the middle and lower parts. However, the center heat map in Fig. 3.4(g) produced by the upper part branch clearly indicates three pedestrian locations, which is crucial to detect the severely occluded third pedestrian. These two examples demonstrate the effectiveness of the UMLF branches in enhancing detection quality in the presence of significant heavy occlusion, without jeopardizing other cases.

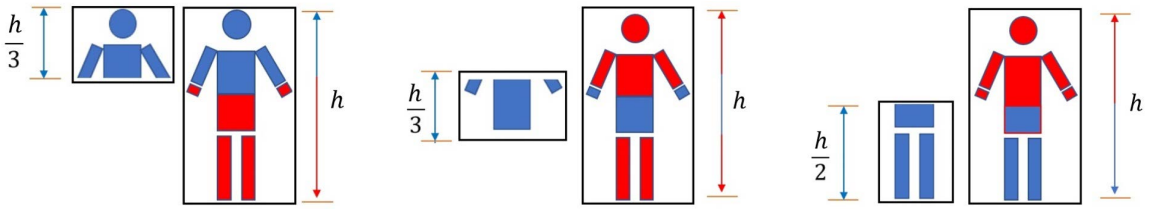
As shown in Fig. 3.3, there is a decoder in each of the four UMLF branches. Each decoder converts the center, scale, and offset maps in each branch to a list of bounding boxes based on their predefined aspect ratios, illustrated in Fig. 3.5. It should be mentioned that a single pedestrian target can be detected multiple times in each of the four UMLF branches, which results in multiple overlapped full-length bounding boxes per branch, creating a type of redundancy referred to as intra-branch redundancy. Moreover, the same pedestrian may be detected by more than one branch, especially if a pedestrian is fully visible in the image, this type of redundancy is referred to as inter-branch redundancy. The post-processing block, presented in the following section, is intended for bounding boxes refinement and redundancy elimination.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 3.4: Two detection examples of the proposed MB-CSP System. (a) (f) Input images. (b) (g) Center heat-maps of the Upper Parts. (c) (h) Center heat-maps of the Middle Parts. (d) (i) Center heat-maps of the Lower Parts. (e) (j) Center heat-maps of the Full Body.




(a) (b) (c)

Figure 3.5: Detection boxes of the three parts and their extension. (a) Upper part. (b) Middle part. (c) Lower part.

## 3.4 Post-Processing Block

The post-processing block is designed to eliminate duplicated pedestrian boxes, and to identify/preserve one bounding box per detected pedestrian. It is performed in two steps to eliminate intra-branch redundancy and inter-branch redundancy, respectively.

For intra-branch redundancy, the duplicated boxes generated in the same branch are removed by means of NMS. It is known that a single pedestrian can be indicated by highly overlapped boxes. The degree of overlapping reflects the likeness of the case, which is measured by IoU index representing an overlap between 0% and 100%. If IoU value of two bounding boxes is higher than a threshold, they are considered to indicate the same pedestrian and the one having the lower confidence score will then be eliminated.

The above-mentioned threshold should be chosen very carefully. As NMS is performed in each of the four branches, the thresholds can be selected differently based on the detection criteria of different body parts. To decrease the risk of false eliminations, the IoU threshold of the upper part is set more cautiously to be 0.6, compared to 0.5 for the other branches. In case of detecting pedestrians that are heavily occluded by other pedestrians, only the upper parts of the occluded pedestrians can be differentiated, while their full-length boxes may highly overlap. In this case, setting the IoU threshold for the upper branch to 0.6 allows to preserve the two individual pedestrians upper parts.

The NMS performed in each of the four branches removes most of intra-branch redundant bounding boxes, and the remaining bounding boxes represent potential pedestrian candidates detected in each branch. The bounding-boxes lists generated by the four branches are then examined together, in the second step, to eliminate inter-branch redundancy.

The inter-branch redundancy can be caused by the detection of a single fully visible, or mostly visible, pedestrian in multiple branches, where the redundant bounding-boxes are usually highly overlapped. However, if two pedestrians are heavily occluded by each other, their boxes generated in the same branch or different branches, can also be overlapped. In order not to falsely eliminate the bounding boxes representing heavily occluded pedestrians, one needs to look into not only the overlap rate, but also other indications from the four bounding boxes lists. The method, referred to as Boosted Identity Aware Non-Maximum Suppression (BIA-NMS), is to check if a group of overlapped boxes represent a single pedestrian or multiple heavily occluded ones.

BIA-NMS is proposed with a view to minimizing the risk of merging heavily overlapped boxes belonging to different pedestrians, while suppressing duplicated pedestrian boxes. The following two points are used to develop BIA-NMS algorithm.

1. BIA-NMS aims at eliminating duplicated detection boxes, generated by different branches, of the same pedestrian target. Hence, no boxes of the same branch can be merged in this procedure, to eliminate the risk of missing occluded targets. To be more specific, at a given location, the boxes to be checked must be from different branches and are eventually merged to be one.

2. At given location, relatively high scores of multiple boxes from different branches indicate a detection of multiple parts of the same pedestrian, implying a high certainty of true detection. In this case, the final detection score will be boosted.

BIA-NMS is performed in the following steps.

1. Sort all the detected boxes in a descending order based on their confidence scores.
2. Identify the box with the highest score and refer to it as $B_{max}$.
3. Calculate the IoU between all the detected boxes and $B_{max}$.
4. Identify the boxes with IoU greater than 0.6 and add them to the new list $B_{duplicate}$, make sure that only one box per branch is added to $B_{duplicate}$ (the box with the highest IoU per branch).
5. Define $N$ as the number of boxes in $B_{duplicate}$.
6. Modify the score of $B_{max}$ as follows:

$$Score_B = (N-1) \times \lambda \times Score_O + Score_O \tag{3.1}$$

   where $Score_B$ denotes the boosted score of $B_{max}$, $Score_O$ is the original score of $B_{max}$ and $\lambda$ is the boosting weight set to 0.08.
7. Add $B_{max}$ and its boosted score to the final detection list.
8. Remove $B_{max}$ and $B_{duplicate}$ from the initial list.
9. Repeat the process starting from step 1.

Fig. 3.6 presents an example of two pedestrians applied to the proposed MB-CSP system. The pedestrian in green is fully visible, hence the UMLF block can detect its upper, middle, lower and full body parts (indicated by the green checkmarks). Meanwhile, the red pedestrian is highly occluded and only the upper part can be detected (red checkmark), and his middle, lower and full body parts are easily missed

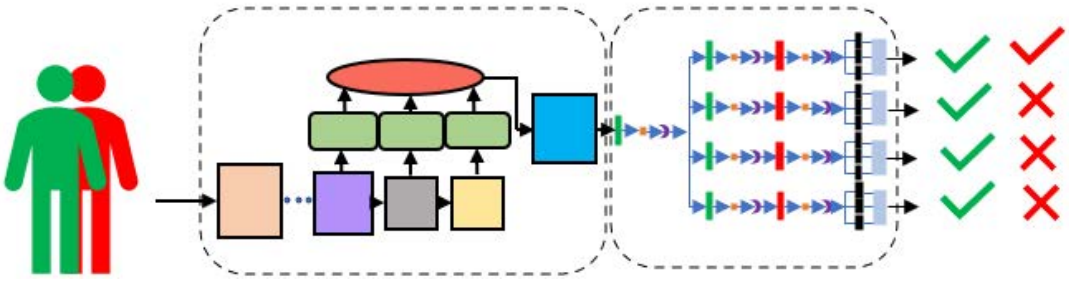
(red-crosses). The five detected boxes are depicted in (b), where the full-box for the green pedestrian is represented by a black colour and has the highest score of 0.8. The process of elimination starts by considering the IoU between all the detected boxes and the box with the highest score (the black box). In this example, the four boxes have IoU values greater than 0.5 with the black box. However, in (c), BIA-NMS eliminates three of the four highly overlapped boxes and preserves one box (violet box). This is because violet boxes represent upper body boxes, and the black box is highly overlapped with two violet boxes, hence only one of them is eliminated (the one with highest IoU value). Finally, the score of the black box is boosted to become 0.99 using equation 3.1.

In Fig. 3.7(a), an image including three pedestrians with different degrees of occlusion is illustrated. If NMS is applied in the second post-processing stage, one of the three pedestrians will be missed in the detection due to the heavy occlusion, as shown in Fig. 3.7(b). The proposed BIA-NMS helps to capture the missed one, so that all the three pedestrians are detected. Fig. 3.7(c) illustrates the detection result, indicated by the three boxes, before the boosting. The scores of the detected pedestrian boxes are boosted, by means of the calculation defined by Equation 3.1, as shown in Fig. 3.7(d).

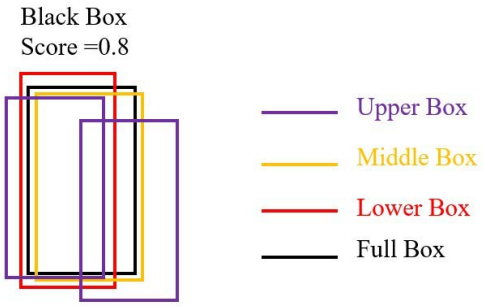
## 3.5 Parts Annotation

Most pedestrians' datasets provide annotation information that specify two bounding boxes for every pedestrian. Visible bounding box that indicates visible area of a pedestrian, and Full bounding box that describes the full pedestrian body including its extension if it is occluded. Annotation information is provided as follows:

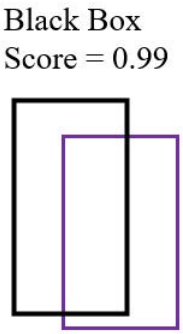
$$Annotation = [x_f, y_f, w_f, h_f, x_v, y_v, w_v, h_v] \tag{3.2}$$



(a)

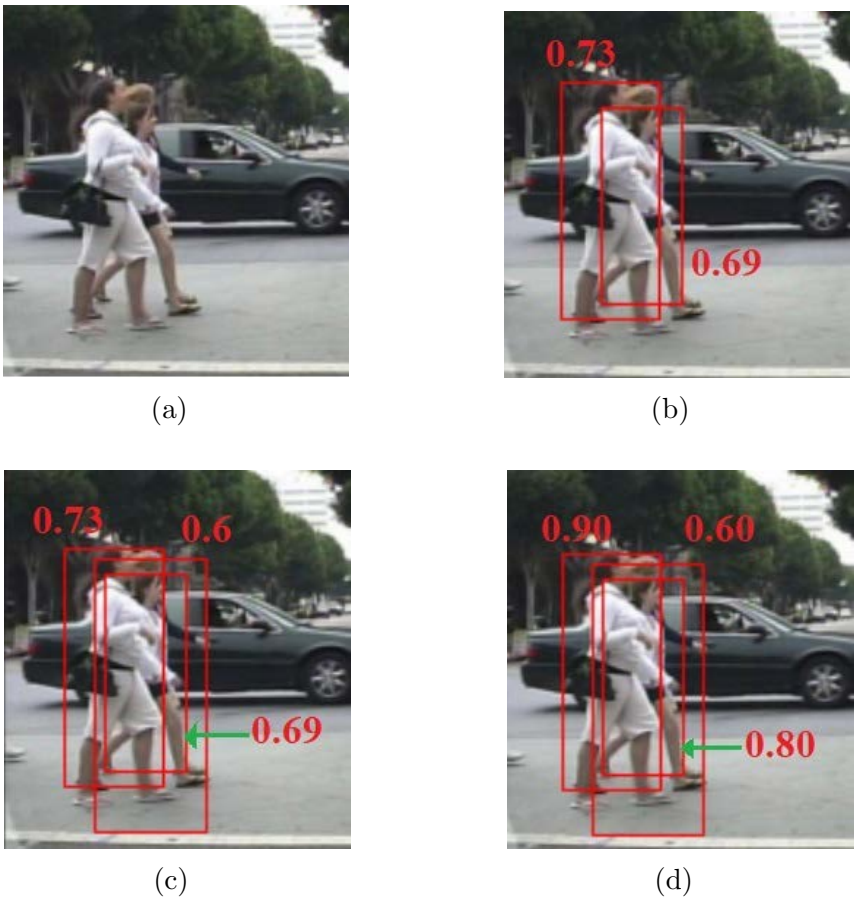



(b)




(c)

Figure 3.6: (a) Example of input involving two pedestrians, of whom one is severely occluded. (b) Boxes generated by the four branches around the pedestrians locations. (c) Result produced by the proposed BIA-NMS. The two boxes from the Upper part branch should not be merged.




(a) (b) (c) (d)

Figure 3.7: (a) Input Image. (b) Detection results by MB-CSP and NMS, (c) by MB-CSP and BIA-NMS before boosting, and (d) by MB-CSP and BIA-NMS after boosting.

where $x_f, y_f$ and $x_v, y_v$ are the coordinates of the top left corner of the full box and the visible box, respectively. $w_f, h_f$ and $w_v, h_v$ are their corresponding width and height.

Since the proposed system has four detection ends, each pedestrian in the image is assigned four bounding boxes, namely $BB_u, BB_m, BB_l$ and $BB_f$, to describe the upper, middle, lower and full pedestrian parts, respectively. Algorithm 1 presents detailed explanation of the annotation algorithm.

## 3.6 System Loss

To calculate the total loss $Loss_T$ of the proposed MB-CSP system, the four branch losses are combined as follows:

$$Loss_T = \alpha_1 Loss_U + \alpha_2 Loss_M + \alpha_3 Loss_L + \alpha_4 Loss_F \tag{3.3}$$

where $Loss_U$, $Loss_M$, $Loss_L$, and $Loss_F$ indicate the system loss of the upper, middle, lower, and full branches, respectively. For simplicity, $\alpha_1$, $\alpha_2$, $\alpha_3$, and $\alpha_4$ are set to 1s. However, adapting different weights can be investigated.

Furthermore, for branch $i$, the branch loss $Loss_i$ can be expressed as:

$$Loss_i = Loss_{C\_i} + Loss_{S\_i} + Loss_{O\_i} \tag{3.4}$$

where $Loss_{C\_i}$, $Loss_{S\_i}$, and $Loss_{O\_i}$ are the center, scale, and offset losses for branch $i$, respectively.

To calculate the center loss for every branch, the same procedure presented in [23] is followed. The main difference is that centers are calculated for every specific part instead of a single center for the entire pedestrian body. Following this procedure, the cross-entropy center loss is defined as:

**Algorithm 1** Parts Annotation.

**Input:**

$BB_f$=$[x_f, y_f, w_f, h_f]$
$BB_v$ =$[x_v, y_v, w_v, h_v]$

**Output:**

$BB_u$=$[x_u, y_u, w_u, h_u]$
$BB_m$ =$[x_m, y_m, w_m, h_m]$
$BB_l$=$[x_l, y_l, w_l, h_l]$
$BB_f$=$[x_f, y_f, w_f, h_f]$

**procedure** PARTS ANNOTATION($BB_{full}, BB_{vis}$)
  **for** $img$ in **Images do**
    **for** $ped$ in **Pedestrians do**
      $BB_u = \left[x_f, y_f, w_f, \frac{h_f}{3}\right]$
      $BB_m = \left[x_f, y_f + \frac{h_f}{3}, w_f, \frac{h_f}{3}\right]$
      $BB_l = \left[x_f, y_f + \frac{h_f}{2}, w_f, \frac{h_f}{2}\right]$
      **if** $\frac{Area(BB_u \cap BB_v)}{Area(BB_u)} > 0.2$ **then**
        $BB_u \leftarrow BB_u$
      **else**
        $BB_u \leftarrow [0, 0, 0, 0]$
      **if** $\frac{Area(BB_m \cap BB_v)}{Area(BB_m)} > 0.2$ **then**
        $BB_m \leftarrow BB_m$
      **else**
        $BB_m \leftarrow [0, 0, 0, 0]$
      **if** $\frac{Area(BB_l \cap BB_v)}{Area(BB_l)} > 0.2$ **then**
        $BB_l \leftarrow BB_l$
      **else**
        $BB_l \leftarrow [0, 0, 0, 0]$
      **Return** $BB_u, BB_m, BB_l, BB_f$

$$Loss_C = \begin{cases} \frac{-1}{N_{obj}} \sum_{i=1}^{\frac{W}{r}} \sum_{j=1}^{\frac{H}{r}} (1-p_{ij})^{\gamma} log(p_{ij}), & y_{ij} = 1 \\ \frac{-1}{N_{obj}} \sum_{i=1}^{\frac{W}{r}} \sum_{j=1}^{\frac{H}{r}} (1-M_{ij})^{\beta} p_{ij}^{\gamma} log(1-p_{ij}), & y_{ij} = 0 \end{cases} \tag{3.5}$$

where $N_{obj}$ is the number of objects (specific body part) in the image, $H$, $W$ and $r$ are the height, width and down-sampling factor of the image, respectively. $M_{ij}$ is a 2d Gaussian map built around the center of every part, based on the height and width of the specific part, this is done to reduce the uncertainty created by the negatives surrounding center points, by reducing their effect on the total loss [23]. $p_{ij}$ is the predicated probability for a center to be presented at location $i, j$, and $y_{ij}$ is the ground truth value, equals to 1 if there is a center at location $i, j$ and 0 otherwise. $\gamma$ and $\beta$ are hyper-parameters, $\gamma$ is set to 2 as recommended by [55], and $\beta$ is set to 4 [58]. Scale and offset losses of every branch are calculated using smooth L1 loss equation as follows:

$$\text{Loss} = \frac{1}{N_{obj}} \sum_{n=1}^{N_{obj}} \text{SmoothL1}(s_n, t_n) \tag{3.6}$$

where $N_{obj}$ is the number of objects (specific body part) in the image, $s_n$ and $t_n$ represent the network's prediction and the ground truth of each positive target, respectively.

## 3.7 Summary

In this chapter, a novel CNN system referred to as MB-CSP is introduced. This system is specifically designed to enhance occluded pedestrian detection. The classification stage consists of four detection branches, each dedicated to generating heat-maps to indicate the locations of different pedestrians body parts. Special post-processing method refereed to as BIA-NMS is introduced, to merge the detections of system branches and generate the final outputs. The chapter also introduced the algorithm

used for annotating pedestrian parts and details the loss functions crucial for the training process. While the MB-CSP system improves occluded pedestrian detection, identifying smaller pedestrians remains challenging. The next chapter highlights these challenges and introduces architectures to tackle them.

# Chapter 4

# Proposed CNN Architectures for Enhanced Small Pedestrian Detection

Pedestrians in images often vary in size and shape, frequently appearing much smaller compared to the original image dimensions. Detecting these small-scale pedestrians poses a significant challenge in creating a reliable detection system. This chapter addresses this challenge by proposing multiple CNN architectures, all based on the multi-branch CNN system introduced in Chapter 3. Section 4.1 discusses the difficulties of detecting small-scale pedestrians and presents the proposed solutions to improve their detection. In Section 4.2, the architectures designed for detecting pedestrians at various scales are presented. Section 4.3 explains the proposed region selection models and their corresponding architectures for small-scale pedestrian detection. Section 4.4 concludes the chapter and summarizes its main contributions.

## 4.1 Overview

In this chapter, the term small-scale pedestrians refers to pedestrians smaller in size relative to the original image dimensions, with the specific size varying based on the dataset. However, this thesis focuses on enhancing the detection of pedestrians with heights between 30 to 80 pixels. As illustrated in Fig. 4.1, this height range corresponds to the medium scale in the Caltech-USA pedestrian dataset, with images having a 480x640 resolution. Fig. 4.1 emphasizes that pedestrians of this size are commonly encountered, thus indicating the importance of enhancing their detection. Despite the challenges in detecting pedestrians at this scale, the right approaches can yield significant improvements. A primary challenge in detecting these pedestrians lies in accurately capturing their distinct features. This becomes evident in situations like densely populated urban areas or surveillance systems, where precise pedestrian detection is essential. This chapter examines two key observations concerning this challenge, laying the groundwork for proposing solutions based on these observations.

1. Pedestrian features that are associated with large-scale pedestrians differ from those associated with smaller pedestrians. Therefore, developing a model to detect pedestrians across all scales poses a challenge. This complexity originates from the need to account for varied feature sets and sizes, making universal detection more complicated.

2. Smaller pedestrians occupy a limited number of pixels in images, especially considering the consecutive downsizing and pooling processes common in many pre-trained detection models. Such operations can further reduce the clarity and distinctiveness of these smaller features, presenting additional hurdles in achieving precise detection. Moreover, many deep learning models, trained on general object detection datasets, mainly focus on detecting larger objects/pedestrians. As a result, these models often struggle to effectively detect smaller pedestrians.

The gap in their training can lead to inaccuracies in real-world scenarios where detection of diverse sizes is essential.

From these observations, two solutions have been proposed to address the specific issues identified. Their potential integration is also under consideration. The proposed solutions are as follows:

1. To address the dissimilarity between the features of large-scale and small-scale pedestrians, this chapter introduces two distinct architectures. One assigns a separate detection system for each scale, named DualScaleSeparateNet (DSSN). The other integrates an additional branch into the previously suggested MB-CSP system, dedicated to identifying small-scale pedestrians; this design is called DualScaleBranchNet (DSBN).

2. To improve the detection of smaller pedestrians, often overlooked due to insufficient pixel information, a region proposal model is introduced. This model identifies and enlarges regions potentially containing smaller pedestrians to enhance their detection accuracy. The region proposal can be done using two methods. The first method uses heat-maps generated by the proposed MB-CSP model, resulting in an architecture named RegionUpscaleNet-HeatMap (RUN-HM). Alternatively, region proposal can be driven by a designated branch within the MB-CSP model, leading to the architecture termed RegionUpscaleNet-DetectorGuided (RUN-DG).

The remainder of this chapter presents the proposed architectures, including their losses, and post-processing techniques.

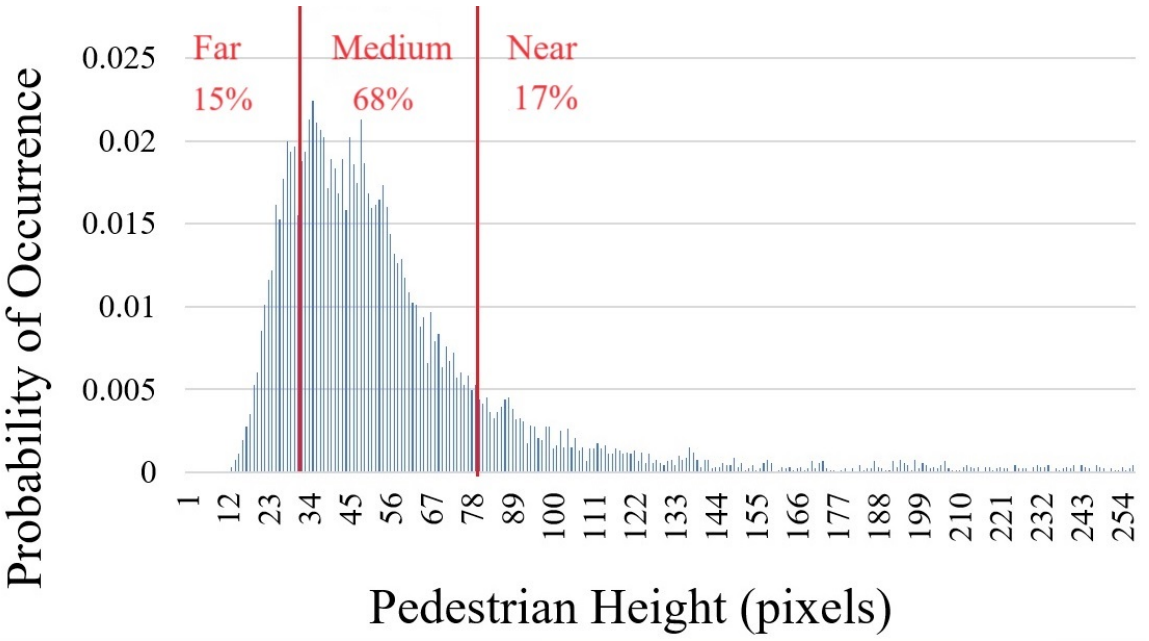



Figure 4.1: Distribution of Pedestrian Heights in the Caltech-USA Dataset.

## 4.2 Architectures to Address Pedestrian Size Variability

In this section, the development of scale-specific systems designed for pedestrian detection is discussed. The detection approach is categorized into two distinct scales.

- **Large Pedestrian Scale**: This scale focuses on pedestrians with larger dimensions, especially those exceeding 80 pixels in height. Pedestrians of this size are often closer to the vehicle, making their accurate detection vital. By omitting smaller, potentially noisy pedestrian samples from the training, the model can converge more effectively and refine its parameters to better identify these larger pedestrians.

- **Small Pedestrian Scale**: This scale focuses on detecting pedestrians with heights ranging from 30 to 80 pixels. While the training emphasizes enhancing the detection of these smaller pedestrians, it might come at the cost of reduced performance in detecting larger ones.

### 4.2.1 Separate Detection Systems for Various Pedestrian Scales

The proposed architecture adopts an ensemble strategy, utilizing two distinct systems. Each system processes input images, yielding individual detection results. To produce the final output, these detections are combined using a specific criterion that eliminates duplicate detections and filters the noisy ones. As illustrated in Fig. 4.2, the proposed architecture is referred to as DualScaleSeparateNet (DSSN) and it integrates a MB-CSP system to detect large-scale pedestrians with a designated CSP [23] system to target small-scale ones. While the two systems setup enhances training accuracy, it demands more computational power for testing because both systems process data simultaneously.

For the DualScaleSeparateNet framework, the two systems have individual loss computations. The large-scale pedestrian system adopts the loss formulation described in equation 3.3. Conversely, the small-scale pedestrian system utilizes the CSP loss model as follows:

$$Loss = Loss_C + Loss_S + Loss_O \tag{4.1}$$

Within this equation, $Loss_C$, $Loss_S$, and $Loss_O$ define the center, scale, and offset losses for the CSP model, respectively.

### 4.2.2 Separate Detection Branches for Various Pedestrian Scales

In a different approach, DualScaleBranchNet (DSBN), as depicted in Fig. 4.3, implements the advantages of diverse detectors in a more simplified manner compared to the DSSN architecture. Instead of operating with completely separate systems, DSBN utilizes a shared Feature Generation Block among all the branches. This architectural choice ensures that the integration of an additional branch requires only the addition

of a few more convolutional layers, making it more resource efficient. This shared structure can potentially lead to more unified feature learning, possibly benefiting the detection of both smaller and larger pedestrians without overly complicating the network architecture. The additional branch designed to detect smaller pedestrians is comprised of the following layers:

1. **Separable Convolutional Layer**: This layer employs a combination of $3 \times 3$ depth-wise convolutions followed by pointwise convolutions. After this convolutional operation, batch normalization is applied, followed by the application of a ReLU activation function.

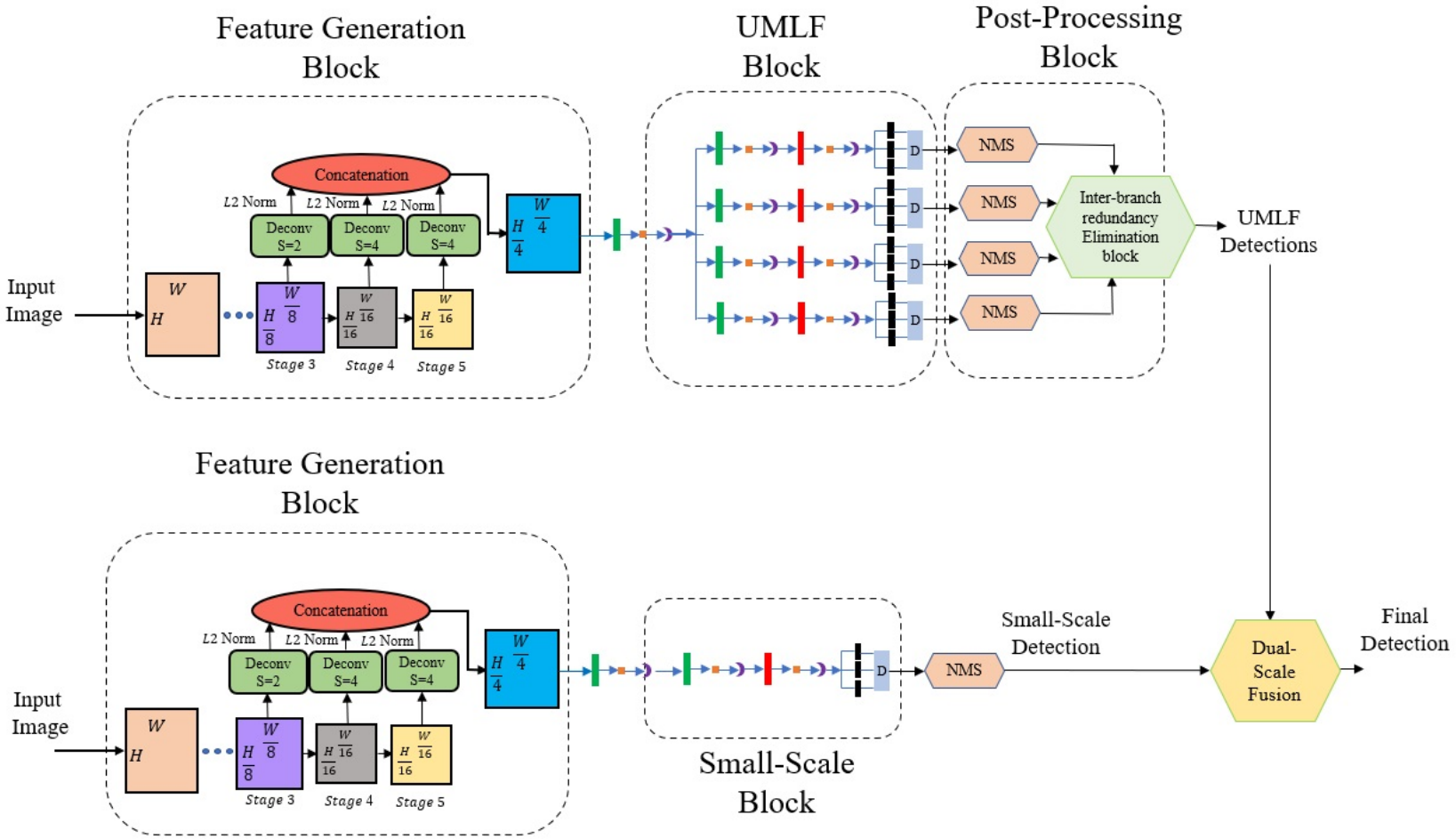
2. **Standard Convolutional Layer**: A conventional convolution is then applied using 256 filters of size $3 \times 3$. This is followed by batch normalization and a subsequent ReLU activation function.

3. **Output Layer**: The outputs from the previous standard convolution are subjected to three distinct $1 \times 1$ standard convolutions. These are employed to generate center heat-map, scale heat-map, and offset heat-maps, especially designed for the branch targeting smaller pedestrians.

The total loss of the DualScaleBranchNet is denoted as $Loss_T$. It is computed by aggregating the individual losses from its five branches, as represented in equation 4.2:

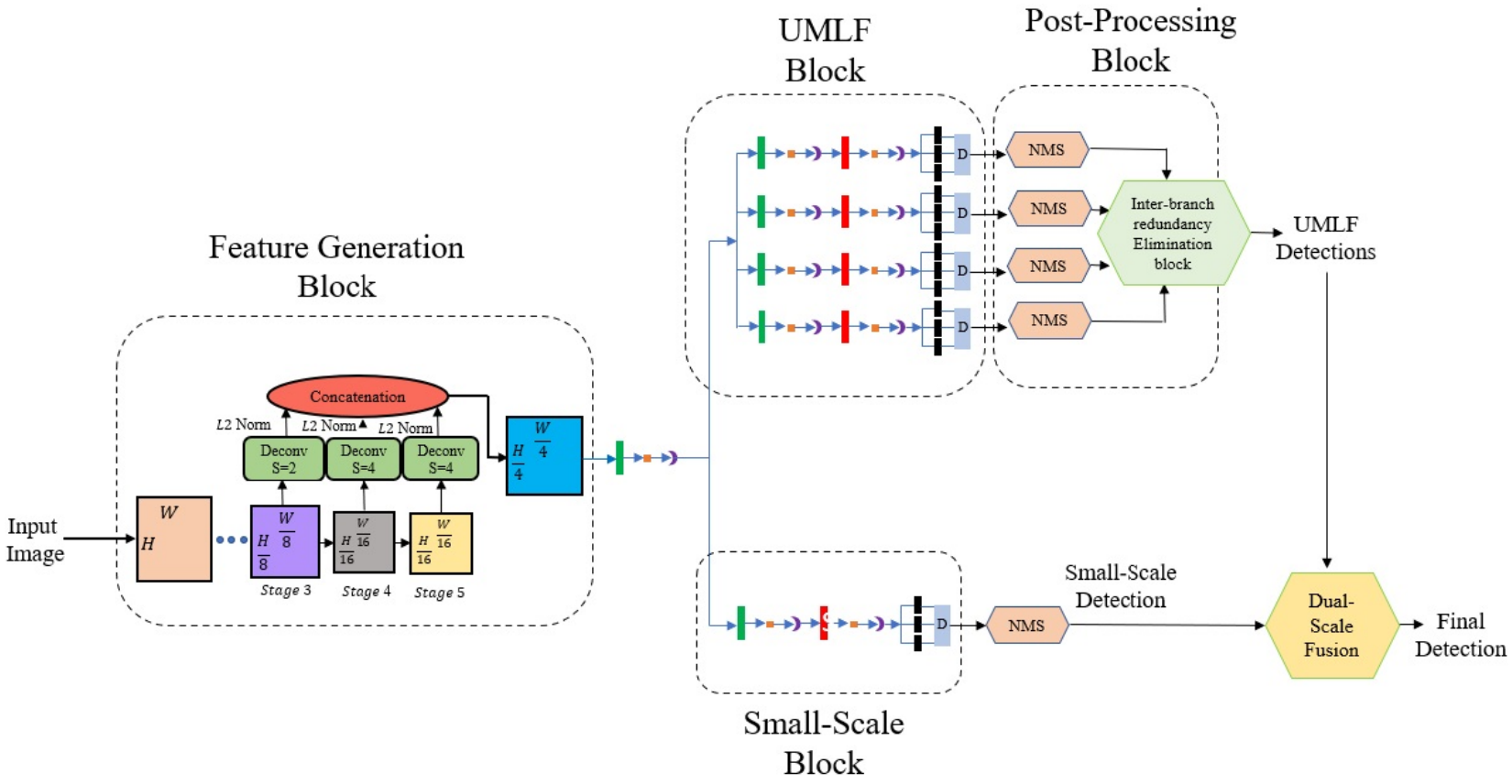
$$Loss_T = \alpha_1 \cdot Loss_U + \alpha_2 \cdot Loss_M + \alpha_3 \cdot Loss_L + \alpha_4 \cdot Loss_F + \alpha_5 \cdot Loss_S \qquad (4.2)$$

Here, $Loss_U$, $Loss_M$, $Loss_L$, and $Loss_F$ denote the upper, middle, lower, and full branches losses, respectively. $Loss_S$ indicates the small branch loss. The coefficients $\alpha_1$ through $\alpha_5$ are set to 1, though their adjustment may optimize the model's performance. Finally, for a given branch, the branch loss is detailed in equation 3.4.




(a)

Figure 4.2: DualScaleSeparateNet Architecture.




(a)

Figure 4.3: DualScaleBranchNet Architecture

Both DualScaleSeparateNet and DualScaleBranchNet architectures have a distinct approach to handling pedestrian detections. For the four branches responsible for upper, middle, lower, and full pedestrians, their detections are fused using the Boosted Identity Aware Non-Maximum Suppression (BIA-NMS), as introduced in Chapter 3. Subsequently, these detections are combined with the detections of the small-scale pedestrian branch in the DualScaleBranchNet system or with the detections of the small-scale pedestrian detector in the DualScaleSeparateNet system using Standard Non-Maximum Suppression (NMS). Notably, if the small-scale detector identifies a pedestrian taller than 80 pixels, its detection is omitted, relying solely on the detections from the other four branches. This strategy ensures that any limitations in the smaller detector's performance for larger pedestrians don't undermine the overall detection accuracy.

## 4.3 Architectures for Enlarging Potential Pedestrian Regions

In this section, another architecture to enhance small-scale pedestrian detection is proposed. The architecture is specifically designed and trained using large-scale pedestrians, ensuring its effectiveness in identifying them during the testing phase. It has been observed that generic architectures, when attempting to detect pedestrians of all sizes, might compromise the accuracy of large-scale pedestrian detection, which is generally more straightforward. Recognizing this challenge, the proposed architecture emphasizes the reliable detection of larger pedestrians by systematically excluding noisy small-scale samples during its training. In its testing procedure, two evaluations are conducted. One on the original image to detect large-scale pedestrians, and another on specific regions chosen for their potential to contain small-scale pedestrians. Once identified, these regions are enlarged, making the smaller pedestrians appear larger

and hence fit within the scale used in training. This strategy enhances the detection of small-scale pedestrians while ensuring the accuracy for larger ones is not compromised.

Figs 4.4a and 4.4b illustrate the distribution of pedestrians' locations. Fig. 4.4a focuses on pedestrians taller than 40 pixels, while Fig. 4.4b highlights those with heights between 30 and 80 pixels. The data shows that smaller pedestrians often appear in the upper sections of the images. To identify the best regions for enlargement within these upper areas, the proposed solution recommends selecting regions where the spatial distribution of small-scale pedestrians, compared to the region's size, matches the distribution of larger pedestrians in the original image. Further details of this distribution comparison method are discussed for Caltech-USA and CityPersons datasets as follows.

**Caltech-USA Dataset Analysis:** In this dataset, pedestrians taller than 40 pixels typically appeared around the $207^{\text{th}}$ row of the image ($\text{row}_{mean1}$), with a standard deviation ($\text{std}_1$) of 53 pixels. Conversely, for pedestrians with heights between 30 and 80 pixels, the central point of distribution ($\text{row}_{mean2}$) was at 182 pixels, with a standard deviation ($\text{std}_2$) of 30 pixels. The objective here is to select image regions and enlarge them, so the pedestrians in the enlarged parts will have the same pedestrian spatial distribution as in the training data of Fig. 4.4a.

$$\text{Ratio}_1 = \frac{\text{std}_1}{\text{H}} = \frac{53}{480} = 0.11 \tag{4.3}$$

By calculating the ratio of $\text{std}_1$ to the full image height (H), the benchmark ratio ($\text{Ratio}_1$) is obtained. This ratio should be matched by $\text{std}_2$ when divided by the height of the selected regions ($\text{H}_{\text{Enlarged}}$) calculated as follows:

$$\text{H}_{\text{Enlarged}} = \frac{\text{std}_2}{\text{Ratio}_1} = \frac{30}{0.11} = 272 \text{ pixels} \tag{4.4}$$


(a) Pedestrian Distribution Across Image Rows For Pedestrians Height Larger than 40 Pixels.


(b) Pedestrian Distribution Across Image Rows for Pedestrians Height Between 30 and 80 Pixels.

Figure 4.4: Pedestrian Distribution in Different Height Categories.

This implies that the optimal height of the selected regions should be 272 pixels. This height is cropped around the mean row so the result is:

$$\text{Starting Pixel} = \text{row}_{\text{mean2}} - \frac{272}{2} = 182 - 136 = 46 \tag{4.5}$$

and

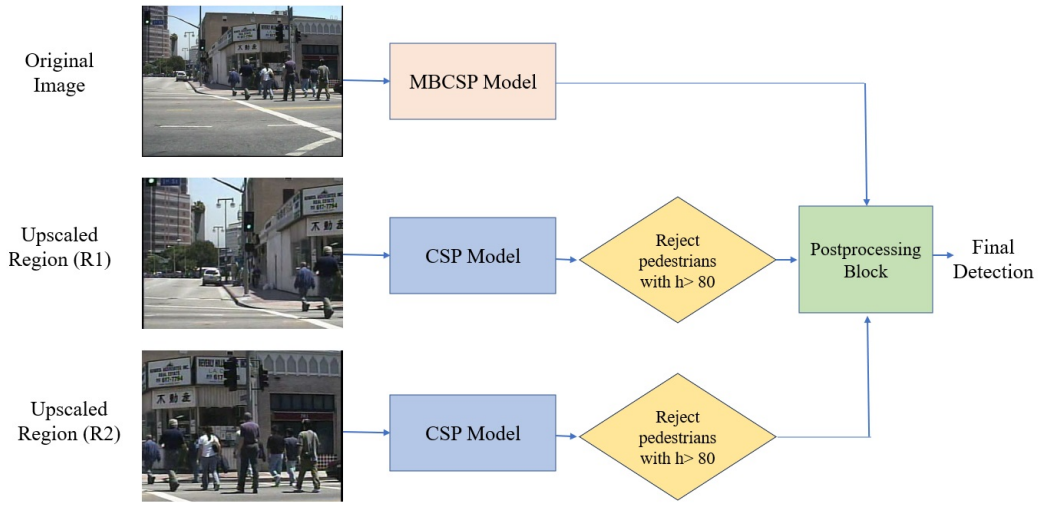
$$\text{Ending Pixel} = \text{row}_{\text{mean2}} + \frac{272}{2} = 182 + 136 = 318 \tag{4.6}$$

To maintain the original image's aspect ratio, it is essential for the selected regions to have a width of 362 pixels. Two regions meeting these specifications are adequate for covering the areas of interest within the image.

**CityPersons Dataset Analysis:** For the CityPersons dataset, characterized by an image height of 1048 pixels and a width of 2048 pixels, the statistical analysis of pedestrian spatial distributions is detailed as follows: Pedestrians exceeding 40 pixels in height are mainly centred around the 512th row, denoted as ($\text{row}_{mean1}$). The corresponding standard deviation for this distribution, represented by ($\text{std}_1$), is approximately 295 pixels. In contrast, the second distribution, encompassing pedestrians with heights ranging from 30 to 80 pixels, exhibited a mean row referred to as ($\text{row}_{mean2}$) of 315 pixels and a standard deviation ($\text{std}_2$) of 167 pixels. Based on these values, two regions with height and width of 580 and 1160, respectively are selected. Whereas the starting height pixel is 25 and the ending height pixel is 605.

Fig. 4.5 presents the first architecture proposed to enhance small-scale pedestrian detection under this approach. Initially, an image is processed by the MB-CSP system, with larger pedestrians primarily being targeted. Next, two specific regions of the image are selected and enlarged for further analysis. These enlarged regions are then processed by the CSP system to target the smaller pedestrians. This architecture is referred to as the Region-Upscale-Net (RUN). It should be mentioned that small-scale

pedestrians that are also occluded are very difficult to detect, hence these cases are ignored and only the CSP system is applied.




(a)

Figure 4.5: (a) Original input image. (b) Image representation of Region (R1) after upscaling. (c) Image representation of Region (R2) after upscaling.

### 4.3.1 Region Selection Using Heat-Maps

The RUN architecture selects regions for enlargement based on the matching process detailed in the previous section. Yet, some of these regions might not contain any pedestrians. As Fig. 4.6 demonstrates, enlarging such regions is inefficient and can decrease the overall detection speed. To optimize the process, this section introduces a method that selectively identifies regions for enlargement and processing based on the center heat-map generated by the full-body branch of the MB-CSP system. Within the RegionUpscaleNet-HeatMap (RUN-HM) framework, only regions that display clear signs of pedestrians in the heat-map are processed further.

A threshold is introduced to guide this decision-making. The threshold evaluates the signal strength within specific regions in the heat-map. Regions with signal strength surpassing this threshold, indicating a higher likelihood of pedestrian presence, are enlarged for processing. Conversely, regions that do not meet the threshold are


(a) First Input Image.

(b) Second Input Image.


(c) Region (R1) Of The First Image.

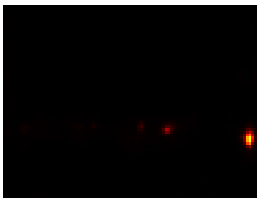
(d) Region (R2) Of The First Image.

(e) Region (R1) Of The Second Image.

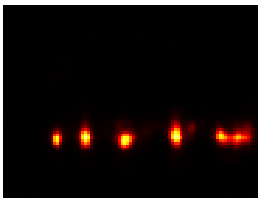
(f) Region (R2) Of The Second Image.

(g) The Center Heat-Map of R1 for the First Image.

(h) The Center Heat-Map of R2 for the First Image.



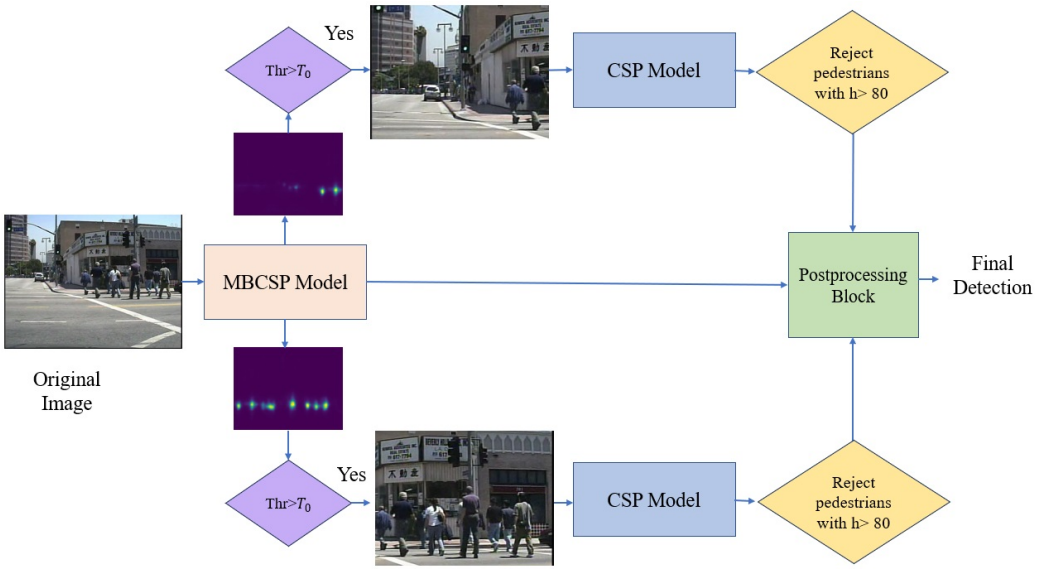
(i) The Center Heat-Map of R1 for the Second Image.



(j) The Center Heat-Map of R2 for the Second Image.

Figure 4.6: Comparison of Two Input Images, Their Highlighted Regions of Interest, and Corresponding Central Heat-Maps.

disregarded. Fig. 4.7 presents the steps and components of the RUN-HM architecture, emphasizing its importance in enhancing the region enlargement and processing strategy. This approach not only optimizes computational resources but also focuses on areas with a greater likelihood of detection, thus improving both the efficiency and accuracy of the pedestrian detection system.



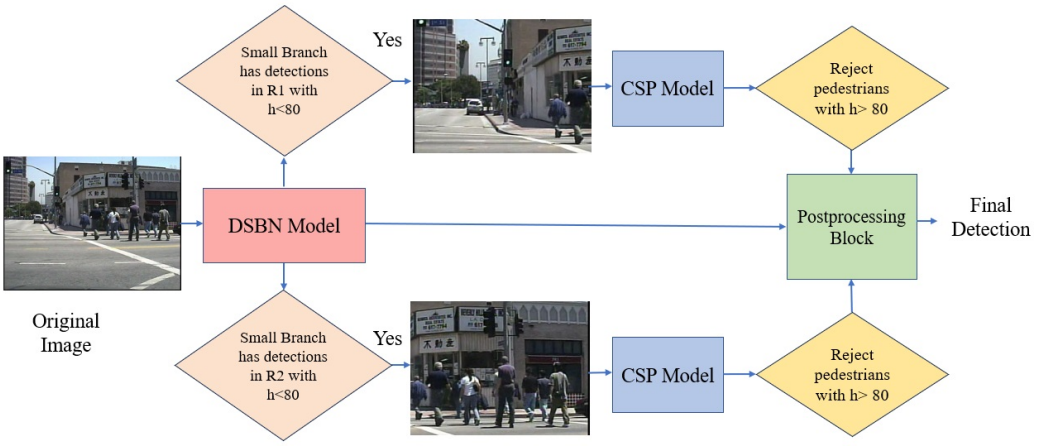

(a)

Figure 4.7: RegionUpscaleNet-HeatMap (RUN-HM) Architecture.

### 4.3.2 Region Selection Using a Dedicated Pedestrian Detector

Similar to RegionUpscaleNet-HeatMap (RUN-HM) detector, the RegionUpscaleNet-DetectorGuided (RUN-DG) focuses on examining regions of interest to determine if they potentially contain pedestrians that require enlargement and further processing. To achieve this, the DualScaleBranchNet (DSBN) system is employed as the foundational system. Specifically, the detections made by the smaller pedestrian branch of the DSBN architecture are assessed. If these detections indicate the presence of pedestrians with heights less than 80 pixels, the corresponding regions are enlarged and processed by the CSP system for further evaluation.

It's important to note that unlike the heat-maps approach used in RegionUpscaleNet-HeatMap (RUN-HM) architecture. The detection guided approach used in RegionUpscaleNet-DetectorGuided (RUN-DG) architecture is more accurate, since this guided detector is specifically trained to detect smaller pedestrians, hence it provides more precise signals, leading to a better choice of regions for enlargement. However, because the foundational detector of RUN-DG architecture is built on the DSBN network, every input image is analysed across all five detection branches, regardless of whether smaller pedestrians are present or not. Fig. 4.8 shows the RUN-DG architecture.




(a)

Figure 4.8: RegionUpscaleNet-DetectorGuided (RUN-DG) Architecture.

For all the proposed architectures discussed in this section, including RUN, RUN-HM, and RUN-DG, the post-processing applied to the detections from the enlarged regions is consistent. Specifically, any detections indicating pedestrians taller than 80 pixels are disregarded. This ensures that the focus remains on detecting pedestrians of smaller sizes within the enlarged regions. Conversely, taller pedestrians are detected using the original image directly, optimizing the detection process for various pedestrian sizes.

## 4.4 Summary

In this chapter, approaches to enhance small-scale pedestrian detection are presented. The initial approach focused on detecting large-scale and small-scale pedestrians separately. This could be achieved either by using two distinct detection systems that process input images independently or by introducing a simplified method of adding an additional branch to the primary large-scale system. This additional branch aims to detect small-scale pedestrians, leading to unified feature learning.

The second approach enhances detection of small-scale pedestrians by enlarging certain image regions likely to contain them. These enlarged regions are then processed using a separate detection system. The aim is to ensure that the spatial distribution of the small-scale pedestrians within these regions matches that of the large-scale pedestrians in the original image. For a more refined region selection, two methods are employed. The first uses heat-map analysis to identify regions, focusing only on areas with clear signs of pedestrians. The second method utilizes a detector specifically designed for smaller pedestrians, leading to more accurate region selection. Overall, these methods tackle the issue of varying pedestrian scales, improving the detection accuracy and simplifying the system computations.

# Chapter 5

# Performance Evaluation

The proposed systems to enhance the detection of occluded pedestrians, as well as systems targeting small-scale pedestrians, have been trained and tested for performance evaluation. This chapter is dedicated to the presentation of their test results. In Section 5.1, the datasets used for the training and testing of the proposed systems are described. Section 5.2 details the experimental setup, explaining the essential parameters and configurations for the systems' implementation. In Section 5.3, the evaluation results of the proposed system for occluded pedestrian detection are presented. Section 5.4 examines the performance of the systems proposed for detecting small-scale pedestrians. Finally, Section 5.5 summarizes the chapter's primary findings and conclusions.

## 5.1 Datasets

This section outlines the two datasets used to assess the proposed systems: Caltech-USA and CityPersons. These datasets are recognized as standard benchmarks for pedestrian detection, presenting a wide variety of pedestrian appearances, such as different poses, occlusions, and sizes.

### 5.1.1 Caltech-USA

Image samples of this dataset are extracted from an approximately 10 hours video recorded by a car driving in the greater Los Angeles area. Images are of size 640 $\times$ 480 pixels. The dataset contains a total of 350,000 labeled bounding boxes in 250,000 frames. For the experiments, one image has been taken out of every 30 frames from the original sequence, resulting in 4250 training images and 4024 testing images. Furthermore, the improved annotation, presented in [59], is adopted for the training and testing. The proposed systems have been evaluated using a log-average miss rate denoted as $MR^{-2}$ for false positive per image in the range ($10^{-2}$ to 1).

The evaluation subsets are presented in Table 5.1. The visibility criteria are based on the visible proportion of a pedestrian, and the height criteria are measured in pixels. Different categories are defined by combining these two metrics to address various scenarios. For example, Bare represents cases where pedestrians have minimal occlusion, whereas Heavy Occlusion involves situations in which pedestrians are significantly obscured. This categorization plays a crucial role in assessing the performance of pedestrian detectors across a range of diverse and challenging scenarios. It ensures a thorough evaluation of their effectiveness under various conditions, covering different degrees of visibility and pedestrian sizes.

Table 5.1: Evaluation Subsets for Caltech-USA and CityPersons Datasets.

| Category | Visibility | Height |
|---|---|---|
| Bare (B) | $\geq$90% | $\geq$50 pixels |
| Reasonable (R) | $\geq$65% | $\geq$50 pixels |
| Medium (M) | 100% | 30 $\leq$ pixels $\leq$ 80 |
| Partial Occlusion (P) | 65% to 90% | $\geq$50 pixels |
| Heavy Occlusion (H) | 20% to 65% | $\geq$50 pixels |

### 5.1.2 CityPersons

This dataset consists of 2975 training images, 500 validation images and 1575 testing images captured in 27 different cities in Germany and neighboring countries. All images are of size 2048× 1024 pixels. The dataset has around 20K pedestrians, where only less than 30% of them are fully visible. The great variation in pedestrian scales, occlusions and backgrounds makes CityPersons a challenging dataset for pedestrian detection. In this thesis, the validation images are used for testing.

## 5.2 Experiments Settings

Simulations have been performed using NVIDIA V100 Volta GPUs with 64G memory. Following the training implementation in [23], the backbone networks are pre-trained on ImageNet, and the total systems are fine-tuned using Adam optimizer. Furthermore, training images have been resized to reduce the training computational complexity. However, the full image size is used in the testing stage. Furthermore, the implementation details are presented in Table 5.2.

Table 5.2: Training Details.

| Dataset | GPUs | Images per GPU | Resized Image | Learning rate | Number of Iterations |
|---|---|---|---|---|---|
| Caltech-USA | 2 | 8 | 448× 336 | $10^{-4}$ | 15K |
| CityPersons | 4 | 2 | 1280× 640 | $2\times10^{-5}$ | 37.5K |

## 5.3 The Results of the Proposed System Targeting Occluded Pedestrian Detection

This section presents the experimental results obtained with the proposed MB-CSP system, specifically targeting the occlusion problem. To begin with, an ablation study is conducted to demonstrate the impact of using different branches during the development of the MB-CSP system. Next, the performance of the proposed MB-CSP is compared with state-of-the-art detectors on the Caltech-USA and CityPersons datasets, respectively.

### 5.3.1 Ablation Study

The proposed MB-CSP system is designed to use the information of the upper, middle, lower and full-body parts in an optimized manner, in order to minimize the interference of the features belonging to the occluding barriers. In this section, three alternatives of UMLF model are investigated, namely, *Upper and Full body parts* (UF) model, *Upper, Middle and Full body parts* (UMF) model and *Upper, Middle and Lower parts* (UML) model. Extensive simulations have been conducted in order to recognize and compare the pros and cons of each model.

UF model is the simplest block to design MB-CSP detector, in which, only upper body box and full pedestrian box are considered. UF model reported the best results compared to other models when tested on heavily occluded pedestrians with a miss rate of 46.62%, as it is clear in Table 5.3. This is expected because lower and middle parts boxes carry no pedestrian information in this case. However, UF performs poorly for the remaining testing subsets.

On the other hand, UMF model, achieved better accuracy compared to UF model on *Reasonable*, *Partial* and *Bare* subsets with miss-rates of 10.35%, 9.64% and 6.74%, respectively. These results indicate the importance of middle body information

for detecting visible and partially occluded pedestrians. Finally, UML utilizes the information in different body parts and neglects full box information. Comparing UML to UMLF model shows a drop in the detection accuracy for most testing subsets when using UML. This observation suggests the importance of the full box information for accurate pedestrian detection.

Table 5.3 compares the processing times of different models for a single image to identify pedestrian locations. The UF model is the fastest at 0.4 seconds per image, while both UMF and UML models take 0.44 seconds. The UMLF model, despite having additional convolutional layers to predict various body parts, only takes slightly longer at 0.48 seconds per image.

Table 5.3: Performance Comparison of Models Incorporating Different Body Part Information: UF (Upper and Full body parts), UMF (Upper, Middle, and Full body parts), UML (Upper, Middle, and Lower parts), and UMLF (all parts). These models aim to optimize body part information use to minimize interference from occluding barriers.

| Method | R | H | P | B | Test-Time |
|---|---|---|---|---|---|
| UF | 12.6% | 46.62% | 11.32% | 8.87% | 0.40 s/img |
| UMF | 10.35% | 46.82% | 9.64% | 6.74% | 0.44 s/img |
| UML | 10.71% | 47.12% | 10.35% | 6.95% | 0.44 s/img |
| UMLF | 10.08% | 47.29% | 10.22% | 6.12% | 0.48 s/img |

### 5.3.2 Comparison with State-of-the-Art Detectors in the Occlusion Challenge

The proposed MB-CSP system has been compared to the state-of-art detectors on Caltech-USA testing sets. MB-CSP refers to the proposed system trained on Caltech-USA training sets, and MB-CSP (City) indicates the system pre-trained on CityPersons training sets and fine-tuned on Caltech-USA training sets. Fig. 5.1 compares the proposed system to the state-of-art detectors reported in Caltech-USA dataset website

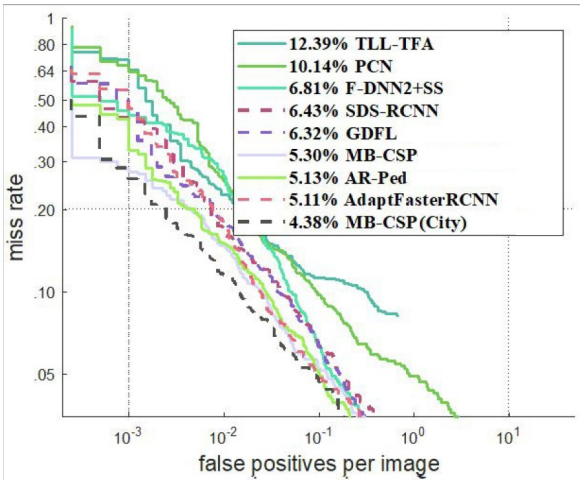
[1]. All the algorithms are evaluated on the improved annotated testing sets, hence there is a variation in their results compared to the ones reported on Caltech-USA website.

In Fig. 5.1 (a), MB-CSP (City) achieved the lowest miss-rate of 4.38% on *Reasonable* subset, Compared to 5.11% for AdaptFasterRCNN [1] and 5.13% for AR-Ped [60]. These results reflect the advantage of using the proposed system for detecting fully visible and partially occluded pedestrians, particularly by boosting pedestrians scores using BIA-NMS method in post-processing. For *Heavy* occlusion subset depicted in Fig. 5.1 (b). MB-CSP (City) and MB-CSP reported superior miss-rates of 27.83% and 30.55%, respectively. Lower by 4.4% compared to the best reported method F-DNN2+SS [61] with a miss-rate of 32.28%. This gain in performance is attributed to the proper design of the multi-branch system.

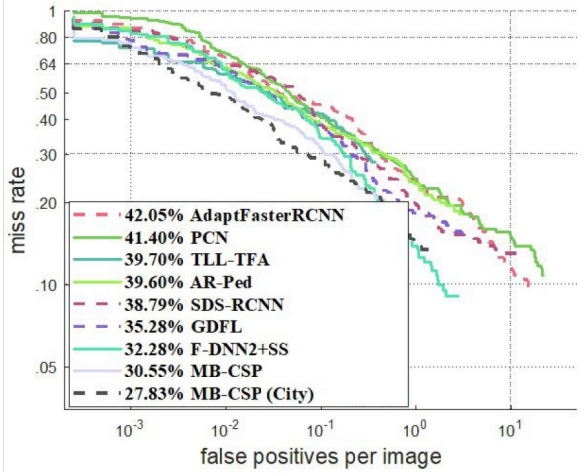
To further investigate the performance of the proposed system. Table 5.4 presents the results of recent state-of-arts detectors that have not been included in Caltech-USA website. The proposed system shows improvement over the Original CSP [23] in all testing subsets. Furthermore, MB-CSP surpassed all detectors in *Reasonable* and *Heavy* occlusion subsets.

The performance of the proposed system is compared to the state-of-the-art methods on CityPerons validation set in Table 5.5. The proposed system in this case, has been trained on CityPersons Dataset. MB-CSP outperformed all the reported methods at all testing subsets. For *Reasonable* and *Bare* subsets, MB-CSP reported miss-rates of 10.08% and 6.12%, respectively. Surpassing the best reported miss-rate by almost 1%. This improvement emphasizes the benefits of using the proposed system in detecting highly visible pedestrians. Furthermore, when detecting occluded pedestrians, MB-CSP scored 47.29% and 10.22% for *Heavy* and *Partial* occlusions, compared to 49.3% and 10.4% for CSP [23]. Proving the superiority of the proposed system in

[1]http://www.vision.caltech.edu/Image_Datasets/CaltechPedestrians/




(a) Reasonable case




(b) Heavy case

Figure 5.1: Comparison of the Proposed MB-CSP System and the State-of-the-Art Detectors on Caltech-USA, Using Average Miss Rate (MR%) on (a) *Reasonable*, and (b) *Heavy* Subsets.

detecting heavily occluded pedestrians with more than 2% gain on Caltech-USA and CityPersons dataset.

Table 5.4: Comparison of the Proposed Multi-Branch Model with State-of-the-Art Methods on the Caltech-USA Dataset.

| Method | R | H |
|---|---|---|
| PAMS-FCN [62] | N.A. | 47.4% |
| CSP [23] | 4.5% | 45.8% |
| CircleNet [63] | 10.2% | 44.5% |
| CSP (City) [23] | 3.8% | 38.5 % |
| FRCN+A+DT [64] | 8.0% | 37.9% |
| Couple [65] | 4.7% | 34.6% |
| MB-CSP | 5.30% | 30.55% |
| MB-CSP (City) | 4.38% | 27.83% |

Table 5.5: Comparison of the Proposed Multi-Branch Model with State-of-the-Art Methods on the CityPersons Dataset.

| Method | R | H | P | B |
|---|---|---|---|---|
| TLL [66] | 14.4% | 52.0% | 15.9% | 9.2% |
| RepLoss [34] | 13.2% | 56.9% | 16.8% | 7.6% |
| OR-CNN [35] | 12.8% | 55.7% | 15.3% | 6.7% |
| Couple [65] | 12.2% | 49.8% | N.A. | N.A. |
| ALFNet [67] | 12.0% | 51.9% | 11.4% | 8.4% |
| CircleNet [63] | 11.7% | 50.2% | 12.2% | 7.1% |
| CSP [23] | 11.0% | 49.3% | 10.4% | 7.3% |
| MB-CSP | 10.08% | 47.29% | 10.22% | 6.12% |

## 5.4 The Results of the Proposed Architectures Targeting Small-Scale Pedestrian Detection

This section presents simulation results for detecting small-scale pedestrians based on the medium category of the Caltech-USA and CityPersons datasets. The proposed

architectures are evaluated for their miss-rate and number of FLOPs, and then compared with state-of-the-art detectors.

### 5.4.1 Assessing Detection Accuracy with Varied Training Height Thresholds

In the original examination of the CSP system as mentioned in [23], the primary focus was on detecting pedestrians classified under the *Reasonable* category. Pedestrians within this category are generally closer to vehicles, necessitating their prompt detection for safety considerations. Consequently, the training primarily included pedestrians taller than 40 pixels. This strategy effectively identified larger pedestrians but encountered difficulties with the *Medium* category, which includes smaller pedestrians. To enhance detection capabilities across different pedestrian scales, the system was revised. This modified version, labeled as $\text{CSP}_{\text{thr20}}$, was trained with a height threshold set at 20 pixels, thus including a wider range of pedestrian heights. However, by including smaller and potentially noisier samples during training, the $\text{CSP}_{\text{thr20}}$ system resulted in reduced detection accuracy for larger pedestrians.

Table 5.6 shows that original CSP system had a 4.5% miss-rate in the *Reasonable* category, while the $\text{CSP}_{\text{thr20}}$ scored 8.18%. However, in the *Medium* category, the $\text{CSP}_{\text{thr20}}$ performed better with a 39.6% miss-rate, compared to the original's 45.97%. These results highlight the fundamental challenge in designing a system that effectively detects both large-scale and small-scale pedestrians.

Table 5.6: Comparison of Pedestrian Detection Accuracy for Two Distinct Height Thresholds in the CSP Systems.

| Method | R | M |
|---|---|---|
| CSP [23] | 4.5% | 45.97% |
| $\text{CSP}_{\text{thr20}}$ | 8.18% | 39.6% |

### 5.4.2 Evaluating Architectures with Separate Detectors for Various Pedestrian Scales

In this experimental section, the results of integrating detections from both small-scale and larger-scale detectors are explored. The architectures under examination employ two distinct strategies. The DualScaleSeparateNet (DSSN) utilizes a separate detector exclusively for identifying small-scale pedestrians, while the DualScaleBranchNet (DSBN) incorporates an additional branch within its framework to achieve the same task. Both architectures, introduced in Chapter 4, aim to enhance pedestrian detection across different scales. The evaluation results of these architectures on the Caltech-USA and CityPersons datasets are presented in Table 5.7.

Based on the analysis of both datasets, the DSSN and DSBN architectures produced comparable results to the original CSP and MB-CSP in the *Reasonable* category. Specifically, on the Caltech-USA dataset, while the MB-CSP recorded a 5.3% miss-rate, both DSSN and DSBN registered miss-rates of 5.5%. This performance similarity suggests that DSSN's additional detector and DSBN's extra branch did not affect the detection efficiency for larger pedestrians, especially with careful post-processing.

For the *Medium* category in the same dataset, the CSP system registered a 45.97% miss-rate. In contrast, the DSSN and DSBN architectures achieved better miss-rates of 37.5% and 37.6%, respectively, highlighting their enhanced capability to detect smaller pedestrians. This improved performance from DSSN and DSBN was expected, given their design to specifically address this scale. Table 5.6 illustrates how previous approaches favoured detection of either large or small pedestrians, highlighting the advantage of the DSSN and DSBN architectures in achieving a balanced detection across both scales.

Table 5.7: Performance Comparison of the Scale-Specific Systems on Caltech-USA and CityPersons Dataset Using Average Miss-Rate ($MR^{-2}$). The DualScaleSeparateNet (DSSN) utilizes a separate detector for small-scale pedestrians, while the DualScale-BranchNet (DSBN) introduces an additional branch for the same purpose.

| Method | Dataset | R | M | H | B | P |
|---|---|---|---|---|---|---|
| CSP [23] | Caltech | 4.5% | 45.97% | 45.8% | N.A. | N.A. |
| MB-CSP | Caltech | 5.3% | 43.39% | 30.55% | N.A. | N.A. |
| DSSN | Caltech | 5.5% | 37.5% | 29.8% | N.A. | N.A. |
| DSBN | Caltech | 5.5% | 37.6% | 33.5% | N.A. | N.A. |
| CSP [23] | City | 11.0% | 32.4 % | 49.3% | 7.3 % | 10.4% |
| MB-CSP | City | 10.08 % | 31.04 % | 47.29% | 10.2% | 6.1 % |
| DSSN | City | 10.6% | 24.7% | 47.6% | 6.9% | 9.8% |
| DSBN | City | 10.02 % | 25.9% | 48.4% | 6.3% | 9.6% |

In the *Heavy* Occlusion category on the Caltech-USA dataset, the DSBN recorded a miss-rate of 33.5%, higher than the 30.55% of the MB-CSP. This difference might be attributed to the introduction of the fifth branch to the DSBN architecture, possibly influencing its training behavior and its capability to detect occluded pedestrians. In contrast, the DSSN showed a 29.8% miss-rate. Using a separate detector for small-scale pedestrians preserved the original MB-CSP detector's efficiency in detecting heavily occluded pedestrians. Evaluation on CityPersons datasets resulted in similar findings as it is clear in Table 5.7.

### 5.4.3 Evaluating Architectures for Enlarging Potential Pedestrian Regions

This section examines the proposed architectures to enhance pedestrian detection by enlarging regions that potentially contain small-scale pedestrians. The proposed architectures are a baseline framework referred to as RegionUpscaleNet (RUN), and its two distinct variants, each designed to a improve the region selection process. One of these variants relies on heat-maps for region selection and is referred to as RegionUpscaleNet-HeatMaps (RUN-HM), while the other utilizes a dedicated detector

to guide this process known as RegionUpscaleNet-DetectorGuided (RUN-DG). These architectures aim to enhance the detection accuracy, especially in challenging scenarios where pedestrians may appear small due to factors like distance or image resolution constraints. Results from evaluations on the Caltech-USA and CityPersons datasets are presented to examine the miss-rate and computational complexity of each architecture.

For both datasets and as indicated in Table 5.8, the baseline RUN architecture and its variants consistently demonstrated improved miss-rates, particularly in the *Medium* and *Reasonable* categories. In the *Medium* category for the CityPersons dataset, the CSP system recorded a miss-rate of 32.46%, the MB-CSP system achieved 31.04%, and the RUN architecture outperformed both with a miss-rate of 24.5%. The variants, *RUN-HM* and *RUN-DG*, performed at 24.58% and 24.47%, respectively. These results can be attributed to the RUN architecture's approach of enlarging regions with high probabilities of containing smaller pedestrians and subsequently processing them through a designated CSP network. Importantly, due to the absence of joint training, there was no observable performance degradation in the *Reasonable* and *Heavy* Occlusion categories.

When evaluating the proposed architectures based on the number of enlarged regions across the Caltech-USA and CityPersons datasets, the RUN architecture consistently exhibits the highest usage, with 8048 enlarged regions for Caltech-USA and 1000 regions for CityPersons. Notably, RUN-HM and RUN-DG show a reduction in enlarged regions, indicating a more selective enlargement strategy. This trend is evident in their respective counts of 3751 and 3214 for Caltech-USA, and 800 and 474 for CityPersons datasets. The main difference between RUN-HM and RUN-DG lies in the criteria for region enlargement. In RUN-HM, regions are enlarged based on the likelihood of heat-map containing any pedestrian activity, rather than exclusively focusing on smaller pedestrians as in RUN-DG. This difference suggests that using RUN-DG could be advantageous in reducing the number of selected regions. However,

it's important to note that the RUN-DG network is built upon the DSBN network, which includes five detection branches, in contrast to RUN-HM, which is based on the MB-CSP network, with just four detection branches. This implies that the fifth branch of RUN-DG, designed for detecting smaller pedestrians, requires extra computations for all images, as it needs to be calculated whenever an image is analysed for pedestrian detection.

Table 5.8: Performance of RegionUpscaleNet Systems on Caltech-USA and CityPersons datasets using Average Miss-Rate (MR%). This table features RUN, a baseline framework that always enlarges regions to improve the small-scale pedestrian detection, and its variants: RUN-HM, which uses heat-maps to select region potentially containing pedestrians, and RUN-DG, driven by a dedicated small-scale pedestrian detector for region selection.

| Method | Dataset | R | M | H | B | Enlarged Regions |
|---|---|---|---|---|---|---|
| CSP [23] | Caltech | 4.5% | 45.97% | 45.8% | N.A | 0 |
| MB-CSP | Caltech | 5.30% | 43.39% | 30.55% | N.A | 0 |
| RUN | Caltech | 4.5% | 34.81% | 30.47% | N.A | 4024+4024 |
| RUN-HM | Caltech | 4.65% | 34.85% | 30.40% | N.A | 1620+2131 |
| RUN-DG | Caltech | 4.95% | 33.58% | 30.25% | N.A | 1468+1746 |
| CSP [23] | City | 11.0% | 32.46 % | 49.3% | 7.3 % | 0 |
| MB-CSP | City | 10.08 % | 31.04 % | 47.29% | 6.12 % | 0 |
| RUN | City | 9.82 % | 24.5% | 47.33% | 6.51% | 500+500 |
| RUN-HM | City | 9.82 % | 24.58% | 47.33% | 6.51% | 386+414 |
| RUN-DG | City | 9.73% | 24.47 % | 47.59 % | 6.49% | 85+389 |

In Table 5.9, the proposed RUN architecture and its variants, RUN-HM and RUN-DG, were benchmarked against the state-of-the-art detectors using the Caltech-USA dataset. Under the *Reasonable* (R) category, both the RUN architecture and the CSP system achieved the best miss-rate of 4.5%. Meanwhile, the RUN-HM and RUN-DG architectures posted slightly higher miss-rates of 4.65% and 4.95%, respectively. In the *Heavy* (H) Occlusion category, the RUN-DG architecture recorded a miss-rate of 30.25%, marking a notable improvement compared to other detectors in the literature, such as GDFL [68] which had a miss-rate of 35.28%. However, the DSSN system emerged as the top performer in this category with a miss-rate of 29.82%. In the

*Medium* (M) occlusion category, RUN-DG achieved the top miss-rate of 33.58%, with the RUN architecture following closely at 34.81%. Notably, the best detector from the literature, GDFL [68], scored 40.26%. These results highlight the advancements in small-scale pedestrian detection achieved by the proposed architectures, without compromising the detection accuracy in other categories.

Table 5.9: Performance Comparison of the Proposed Systems with the State-of-the-Art on Caltech-USA Dataset.

| Method | R | H | M |
|---|---|---|---|
| GDFL [68] | 6.32% | 35.28% | 40.26% |
| TLL-TFA [66] | 12.39% | 39.70% | 44.58% |
| PCN [69] | 10.14% | 41.40% | 54.76% |
| SDS-RCNN [70] | 6.43% | 38.79% | 51.34% |
| CSP [23] | 4.5% | 45.8% | - |
| MB-CSP | 5.30% | 30.55% | 43.39 % |
| DSBN | 5.53% | 33.51% | 37.63% |
| DSSN | 5.59 % | 29.82%% | 37.52% % |
| RUN | 4.5% | 30.47% | 34.81% |
| RUN-HM | 4.65% | 30.40% | 34.85% |
| RUN-DG | 4.95% | 30.25 % | 33.58% |

Table 5.10 presents the computational cost and system size of the various detection systems proposed in this thesis for comparison, with these aspects measured by Floating-Point Operations Per second (FLOPs) and the number of parameters. The CSP system [23] reported the best efficiency, requiring only 192 billion FLOPs and utilizing 40 million parameters. Furthermore, the MB-CSP system shows an increase in computational demand, utilizing 293 billion FLOPs and 42 million parameters. As for the DualScaleBranchNet (DSBN) system, that requires additional branch to detect small-scale pedestrians, the system requires 319 billion FLOPs and requires 43 million parameters. On the other hand, the DualScaleSeparateNet (DSSN), which dedicates a separate stand-alone detector to target small-scale pedestrians presented a large

increase in the computational power and system size at 485 billion FLOPs and 82 million parameters.

For the region selection architectures, the computational demand is typically larger and not constant. The number of times a region is selected for enlargement and further processing primarily depends on whether a specific image indicates the presence of pedestrians. To provide a general perspective, suppose two regions from every image were enlarged and processed as in the benchmark RegionUpscaleNet (RUN) architecture. In this case, the total computational demand would amount to 677 billion FLOPs, and the system would comprise 82 million parameters. These numbers can be significantly reduced if the region selection process uses the proposed heat-maps in the RegionUpscaleNet-HeatMaps (RUN-HM) architecture to signal the presence of pedestrians or by integrating an additional branch in the RegionUpscaleNet-DetectorGuided (RUN-DG) variant to indicate the same. In evaluations, RUN-HM demanded 469 billion FLOPs and 82 million parameters, whereas RUN-DG required 444 million FLOPs and 83 million parameters. Interestingly, even though RUN-DG inherently holds more complexity than RUN-HM, it necessitates fewer computations. This is mainly because RUN-DG is designed to suggest fewer regions for enlargement, a trend particularly evident in the Caltech-USA dataset. However, when both systems enlarge the same number of regions, RUN-HM stands out as the more computationally efficient option. In conclusion, finding a balance between computational demands and system performance is essential, emphasizing the importance of thoughtful design and optimization in real-world applications.

Table 5.10: Comparison of FLOPs and System Parameters Across Different Proposed Systems.

| Method | FLOPs(G) | Parameters(M) |
|---|---|---|
| CSP [23] | 192 | 40 |
| MB-CSP | 293 | 42 |
| DSBN | 319 | 43 |
| DSSN | 485 | 82 |
| RUN | 677 | 82 |
| RUN-HM (Caltech) | 469 | 82 |
| RUN-DG (Caltech) | 444 | 83 |

## 5.5 Summary

In this chapter, the performance of the proposed systems is evaluated using the Caltech-USA [11] and CityPersons [1] datasets. While the MB-CSP system shows significant enhancement in the detection of pedestrians in the *Heavy* Occlusion category, there remains room for advancement in the *Medium* category. In the context of Scale-Specific and Region-Upscale systems, improvements in detecting *Medium* pedestrians are noted, with DSBN being identified as the most computationally efficient. The best results in the detection of *Medium* pedestrians are achieved by the RUN system. However, similar performance levels with reduced computational demands are preserved by its variants, RUN-HM and RUN-DG. Maintaining balance between computational needs and system efficiency is essential, emphasizing the value of careful design of the detection systems in practical applications.

# Chapter 6

# Conclusion

Pedestrian detection is essential for various applications, such as self-driving vehicles, video surveillance, and intelligent street traffic management. However, the wide variations in pedestrian sizes, postures, locations, and backgrounds make their detection a complex task. In particular, the detection becomes significantly challenging due to the lack of pedestrian information when a pedestrian is occluded by objects, such as vehicles or trees, or when they appear at a very small size in the image. Such situations are frequently encountered in the real world. The objective of this thesis has been to design CNN-based pedestrian detection architectures to improve the detection of occluded and small-scale pedestrians.

The first part of this work has addressed the occlusion problem by proposing a specific detection system referred to as Multi-Branch Center and Scale Prediction (MB-CSP). The proposed system employs a multi-branch structure to optimize the utilization of the features extracted from the visible parts of pedestrians. This multi-branch structure enables the feature data from the upper, middle, and lower parts of a pedestrian, as well as those of the full body, to be processed separately. By doing so, the data representing the true pedestrian appearances, whether partially or fully visible, can dominate the final decision. As a result, the interference from

non-pedestrian data in the detection can be minimized. A part annotation algorithm has been introduced to support multi-branch training. Additionally, a new method, termed BIA-NMS, has been developed to optimize the fusion of the detection outcomes from multiple branches. The BIA-NMS method eliminates redundant detections across branches and boosts the scores of the preserved detections.

To improve the detection of small-scale pedestrians, the second part of this work has introduced two approaches, both involving the proposed MB-CSP model. The first approach detects pedestrians of different scales separately by training models to distinguish features unique to each scale category. This is achieved by either integrating an additional detection branch or adding an independent small-scale pedestrian detector, thereby enhancing pedestrian detection across various scales. The second proposed approach identifies regions in the image likely containing small-scale pedestrians and enlarges these regions to enhance their detection. To optimize the region selection process, the model utilizes heat-maps generated by the MB-CSP model or locations suggested by an additional branch dedicated to predict small-scale pedestrian locations.

The detection systems presented in this thesis have been trained and evaluated using images from the Caltech-USA and CityPersons datasets. The results have emphasized the effectiveness of the proposed multi-branch system in detecting occluded pedestrians. Furthermore, testing results have demonstrated that both approaches, specifically designed for small-scale pedestrian detection, significantly improve the accuracy in this category.

# References


[1] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4457–4465.

[2] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International Journal of Computer Vision*, vol. 38, pp. 15–33, 06 2000.

[3] P. Viola and M. Jones, "Robust real-time face detection," in *IEEE International Conference on Computer Vision. (ICCV)*, 2017, pp. 4457–4465.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 886–893.

[5] P. Dollar, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *Proceedings of the British Machine Vision Conference (BMVA)*, 2010, pp. 68.1–68.11.

[6] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *IEEE 12th International Conference on Computer Vision*, 2009, pp. 32–39.

[7] W. Ouyang and X. Wang, "Single-pedestrian detection aided by two-pedestrian detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, 06 2013.

[8] P. Dollár, Z. Tu, P. Perona, and S. J. Belongie, "Integral channel features." in *BMVC*, 2009, pp. 1–11.

[9] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[10] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1751–1760.

[11] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.

[12] R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *Computer Vision (ECCV)*, 2015, pp. 613–627.

[13] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, "From handcrafted to deep features for pedestrian detection: A survey," 2020.

[14] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 82–90.

[15] D. Ribeiro, J. C. Nascimento, A. Bernardino, and G. Carneiro, "Improving the performance of pedestrian detectors using convolutional learning," *Pattern Recognition*, vol. 61, pp. 641–649, 2017.

[16] Y. Zhu, J. Wang, C. Zhao, H. Guo, and H. Lu, "Scale-adaptive deconvolutional regression network for pedestrian detection," 03 2017, pp. 416–430.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[19] R. Girshick, "Fast r-cnn," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, 2015.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[23] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5182–5191.

[24] M. Mathias, R. Benenson, R. Timofte, and L. V. Gool, "Handling occlusions with franken-classifiers," in *IEEE International Conference on Computer Vision*, 2013, pp. 1505–1512.

[25] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[26] Bo Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *IEEE International Conference on Computer Vision (ICCV)*, vol. 1, 2005, pp. 90–97.

[27] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3506–3515.

[28] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *IEEE International Conference on Computer Vision*, 2013, pp. 2056–2063.

[29] M. Shang, D. Xiang, Z. Wang, and E. Zhou, "V2f-net: Explicit decomposition of occluded pedestrian detection," 2021.

[30] J. Noh, S. Lee, B. Kim, and G. Kim, "Improving occlusion and hard negative handling for single-stage pedestrian detectors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 966–974.

[31] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in cnns," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6995–7003.

[32] Z. Guo, W. Liao, Y. Xiao, P. Veelaert, and W. Philips, "Deep learning fusion of rgb and depth images for pedestrian detection," in *British Machine Vision Conference*, 2019.

[33] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 3820–3834, 2020.

[34] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7774–7783.

[35] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware r-cnn: Detecting pedestrians in a crowd," in *Computer Vision (ECCV)*, 2018, pp. 657–674.

[36] S. Liu, D. Huang, and Y. Wang, "Adaptive nms: Refining pedestrian detection in a crowd," 2019.

[37] X. Huang, Z. Ge, Z. Jie, and O. Yoshie, "Nms by representative region: Towards crowded pedestrian detection by proposal pairing," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 747–10 756.

[38] P. Zhou, C. Zhou, P. Peng, J. Du, X. Sun, X. Guo, and F. Huang, "Noh-nms: Improving pedestrian detection by nearby objects hallucination," 2020.

[39] N. O. Salscheider, "Featurenms: Non-maximum suppression by learning feature embeddings," 2020.

[40] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms improving object detection with one line of code," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5562–5570.

[41] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Li, "Occlusion-aware r-cnn: Detecting pedestrians in a crowd," 07 2018.

[42] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 953–961.

[43] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 09 2018.

[44] G. Chen, H. Wang, K. Chen, Z. Li, Z. Song, Y. Liu, W. Chen, and A. Knoll, "A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 2, pp. 936–953, 2022.

[45] L. Cui, R. Ma, P. Lv, X. Jiang, Z. Gao, B. Zhou, and M. Xu, "Mdssd: multi-scale deconvolutional single shot detector for small objects," *Science China Information Sciences*, vol. 63, 02 2020.

[46] Z. Meng, X. Fan, X. Chen, M. Chen, and Y. Tong, "Detecting small signs from large images," 2017.

[47] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2018.

[48] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao, "R-cnn for small object detection," 2017, pp. 214–230.

[49] S. Bell, C. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," 06 2016, pp. 2874–2883.

[50] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1951–1959.

[51] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Sod-mtgan: Small object detection via multi-task generative adversarial network," in *15th European Conference*, 09 2018, pp. 210–226.

[52] Y. Pang, J. Cao, J. Wang, and J. Han, "Jcs-net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 12, pp. 3322–3331, 2019.

[53] C. Wilms and S. Frintrop, *AttentionMask: Attentive, Efficient Object Proposal Generation Focusing on Small Objects*, 06 2019, pp. 678–694.

[54] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," 2019.

[55] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[56] A. Abdelmutalab and C. Wang, "Pedestrian detection using mb-csp model and boosted identity aware non-maximum suppression," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24 454–24 463, 2022.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[58] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints." in *The European Conference on Computer Vision (ECCV)*, 2018.

[59] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1259–1267.

[60] G. Brazil and X. Liu, "Pedestrian detection with autoregressive network phases," in *IEEE Computer Vision and Pattern Recognition*, 06 2019.

[61] X. Du, M. El-Khamy, V. I. Morariu, J. Lee, and L. S. Davis, "Fused deep neural networks for efficient pedestrian detection," *CoRR*, 2018.

[62] P. Yang, G. Zhang, L. Wang, L. Xu, Q. Deng, and M. H. Yang, "A part-aware multi-scale fully convolutional network for pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 1125–1137, 2021.

[63] T. Zhang, Z. Han, H. Xu, B. Zhang, and Q. Ye, "Circlenet: Reciprocating feature adaptation for robust pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4593–4604, 2020.

[64] C. Zhou, M. Yang, and J. Yuan, "Discriminative feature transformation for occluded pedestrian detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 9556–9565.

[65] T. Liu, W. Luo, L. Ma, J. J. Huang, T. Stathaki, and T. Dai, "Coupled network for robust pedestrian detection with gated multi-layer feature extraction and deformable occlusion handling," *IEEE Transactions on Image Processing*, vol. 30, pp. 754–766, 2021.

[66] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on topological line localization and temporal feature aggregation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 09 2018.

[67] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 09 2018.

[68] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for robust pedestrian detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 3820–3834, 2020.

[69] S. Wang, "PCN: part and context information for pedestrian detection with cnns," in *British Machine Vision Conference (BMVC)*, 2017.

[70] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection and segmentation," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4960–4969.