AI

MDPI

Article

Discriminative Deformable Part Model for Pedestrian Detection with Occlusion Handling

Shahzad Siddiqi, Muhammad Faizan Shirazi and Yawar Rehman *

Department of Electronic Engineering, NED University of Engineering and Technology, Karachi 75270, Pakistan; siddiqi@neduet.edu.pk (S.S.); faizanshirazi@neduet.edu.pk (M.F.S.)
* Correspondence: yawar@neduet.edu.pk

Abstract: Efficient pedestrian detection plays an important role in many practical daily life applications, such as autonomous cars, video surveillance, and intelligent driving assistance systems. The main goal of pedestrian detection systems, especially in vehicles, is to prevent accidents. By recognizing pedestrians in real time, these systems can alert drivers or even autonomously apply brakes, minimizing the possibility of collisions. However, occlusion is a major obstacle to pedestrian detection. Pedestrians are typically occluded by trees, street poles, cars, and other pedestrians. State-of-the-art detection methods are based on fully visible or little-occluded pedestrians; hence, their performance declines with increasing occlusion level. To meet this challenge, a pedestrian detector capable of handling occlusion is preferred. To increase the detection accuracy for occluded pedestrians, we propose a new method called the Discriminative Deformable Part Model (DDPM), which uses the concept of breaking human image into deformable parts via machine learning. In existing works, human image breaking into deformable parts has been performed by human intuition. In our novel approach, machine learning is used for deformable objects such as humans, combining the benefits and removing the drawbacks of the previous works. We also propose a new pedestrian dataset based on Eastern clothes to accommodate the detector's evaluation under different intra-class variations of pedestrians. The proposed method achieves a higher detection accuracy on Pascal VOC and VisDrone Detection datasets when compared with other popular detection methods.

Keywords: pedestrian detection; machine learning; occlusion; deformable parts; detection accuracy; intra-class variations; evaluation metric; mean average precision

check for updates

Academic Editor: Giovanni Diraco

Received: 5 February 2025
Revised: 16 March 2025
Accepted: 25 March 2025
Published: 3 April 2025

Citation: Siddiqi, S.; Shirazi, M.F.; Rehman, Y. Discriminative Deformable Part Model for Pedestrian Detection with Occlusion Handling. *AI* **2025**, *6*, 70. https://doi.org/10.3390/ai6040070

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Pedestrian detection has recently attracted a considerable amount of investigation owing to its importance in numerous practical applications such as automatic driving systems, video surveillance, and robotic automation, where efficient and accurate pedestrian detection is critical. Improved detection significantly strengthens road security by helping cars and other vehicles efficiently recognize pedestrians, thereby minimizing the possibility of road accidents. However, occlusion is still a major obstacle for achieving satisfactory detection results. Occlusion occurs when a pedestrian is obscured or blocked from view by other objects in the scene like cars, trees, street poles, or some other pedestrians.

Traditional detection methods assume that pedestrians are fully visible or little occluded; hence, the detection performance of the models falls significantly as the occlusion level of the pedestrian increases. Hence, a more reliable pedestrian detector is needed that can efficiently handle the occlusion issue to obtain more accurate detection results. We

develop a mathematical model to increase the detection accuracy for occluded deformable objects, such as humans, in an image.

To better understand the novelty, it is helpful to contrast the proposed method with previous deformable object detection techniques.

- **Felzenszwalb et al.** [1]: In their work, they introduced a deformable part-based model for human detection that involved manually selecting parts of the human body (e.g., head, torso, arms, etc.). The selection was based on human intuition and each part was treated as a rigid component of the human figure. However, this method lacked optimization of the part selection, as it was reliant on predefined parts, which made it less adaptable to varying human poses and occlusions.
- **Machine learning-based optimization in traffic sign detection** [2]: In contrast to Felzenszwalb et al.'s approach [1], some methods for traffic sign detection used machine learning to optimize the selection of object parts, as seen in [2] for traffic sign detection. Here, the part selection process was automated through learning from data, removing the reliance on human intuition. However, these methods were mainly designed for rigid objects like traffic signs, where the parts did not exhibit significant deformation.
- **Proposed Approach**: The proposed work integrates the best of both works by using machine learning to automatically select deformable parts of the human body, overcoming the limitations of prior methods. The key difference is that it dynamically optimizes part selection based on the actual deformations present in the human body, rather than relying on static, pre-defined rigid parts. This method is more adaptable to the variability in human poses and occlusions.

The novelty of the proposed work lies in its innovative approach to human image detection by utilizing machine learning for selecting deformable parts of a human body. In contrast to previous methods that relied on human intuition for part selection (such as in Felzenszwalb et al. [1] for deformable part models), this method eliminates the need for manual intervention and instead uses machine learning to optimize part selection dynamically.

One of the main distinctive features of pedestrians is that they exhibit a high intra-class variance, which makes the task of detecting them highly challenging.

High intra-class variance refers to the significant variability within the same object class, such as variations in appearance, shape, and pose. Humans specifically wear clothes that may vary geographically from region to region. This affects the basic distinctive features of a human, i.e., contours and edges. Figure 1 shows this effect of intra-class variance in pedestrians. In Figure 1a, the discriminative body parts of the pedestrians are clearly visible and can easily be recognized by the detector, whereas in Figure 1b, the costumes of the pedestrians, which are common in Eastern Muslim culture, are such that their discriminative body features are not clearly visible and hence may cause a detection error.

A detector trained on a dataset containing Western cultural details may fail to recognize humans in Eastern culture. We propose a dataset recorded on Pakistan roads for different apparel to address intra-class variance in pedestrians by capturing the human data with the national costume of eastern culture. Our main contributions are as follows:

- We develop a mathematical model to increase the detection accuracy for occluded deformable objects, such as pedestrians, in an image. In order to accomplish this task, we develop a Discriminative Deformable Part Model (DDPM) by using the concept of breaking a human image into some specified deformable parts.

- We propose a Discriminative Deformable Part Model (DDPM), which uses machine learning to automatically learn the part selection instead of human intuition and incorporates the deformable equations to handle deformable objects such as pedestrians instead of rigid objects, thus yielding a significant increase in detection accuracy for occluded pedestrians. We design a framework for occlusion handling in pedestrians by considering human samples from existing benchmark datasets. Deformable patches from those samples have been automatically selected via machine learning. We develop a set of deformation equations to perform the analysis to relate the deformable body parts of pedestrians.
- To handle contour/edge variation, we propose a dataset recorded on Pakistan roads for different apparel. We capture pedestrian data with the national costume of Pakistani culture (for both male and female costumes), and these data, along with the Western dataset, are used to check and evaluate the proposed technique, and the results of both datasets are analyzed.



Figure 1. Effect of intra-class variance in pedestrians. (**a**) Pedestrians with distinct discriminative body parts. The head, legs, arms, and torso are easily identifiable. (**b**) Pedestrians with Eastern clothing in which discriminative body parts (head, legs, arms, and torso) are hidden and difficult to recognize.

2. Related Works

For pedestrian objects with excellent visibility and a low proportion of occlusion, the majority of pedestrian detectors have acceptable detection capabilities. However, the pedestrian detector's performance may suffer when the pedestrian target is extremely obscured. This section examines some important novel approaches to the occlusion problem that have been put forward by different scholars.

2.1. Traditional Methods and Deep Learning-Based Approaches

Occlusion causes a significant decrease in the detection performance of most state-of-the-art methods. As an example, on the Caltech pedestrian benchmark [3], the pedestrian miss rate is only 19% with zero occlusion, which increases to 39% with occlusion level up to 35% and rises up more remarkably to 78% with occlusion level up to 80%. Occlusion is classified into four groups according to their level of severity [4]: 0% (never occlusion), 1–35% (partial occlusion), 35–80% (heavy occlusion), and above 80% (full occlusion). Traditional pedestrian detection algorithms offer fine detection accuracy under occlusion

percentage between 0 and 10% [5]. But as occlusion level starts to increase, the detection accuracy falls drastically. The algorithm nearly fails to detect the pedestrians where the occlusion is greater than 50%.

Figure 2 illustrates the year-wise gradual evolution of occluded pedestrian detection methods. A. Mohan et al. proposed Haar [6] in 2000, which is considered the foundation of pedestrian detection technology. In this model, the author uses the idea of breaking the human body of the pedestrian into four deformable parts: the head, legs, right arm, and left arm. Although the paper presents a significant contribution to object detection, the method might struggle with detecting objects under varying light conditions, angles, or viewpoint changes.

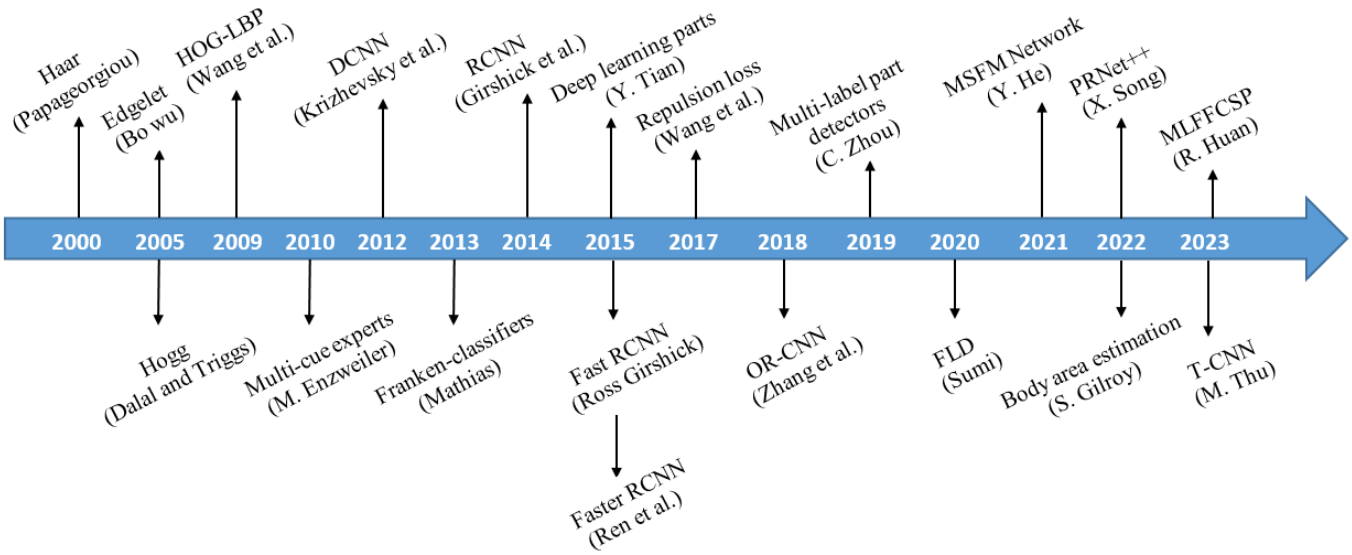

Figure 2. Key development of occluded pedestrian detection [6–25].

Gang Li et al. [26] worked on reducing the number of false positive samples that are generated by the traditional detectors because they treat those pedestrians also as positive samples, which are disregarded during the annotations of the dataset. To address this issue, an additional visible box is used for every bounding box during the detection stage, and those visible boxes are analyzed only during the evaluation stage to match with the disregarded pedestrians. The proposed new evaluation strategy could potentially introduce complexity into the pedestrian detection evaluation process.

2.2. Multi-Cue and Feature Fusion Techniques

Traditional pedestrian detection systems use four stages to handle intra-class variation of pedestrians: feature recognition, distinguishing deformable parts of the body, identification of occluded regions, and classification of the object as a positive or negative candidate. Ref. [27] establishes a connection between these blocks by using Histograms of Oriented Gradients (HOGs) for feature extraction, which are designed to take into account the intra-class variations. The method proposed in the paper involves jointly learning deep features, deformable parts, and occlusion handling, which increases computational complexity. This can result in slower detection times, especially for real-time applications.

Occlusion causes obstruction of some of the body parts of a pedestrian; hence, the corresponding detector havsa greater chance of missing such body parts, thus giving a low overall detection score. Ref. [10] uses the idea to estimate the visibility of the body parts by considering the weights of those components only, which are occlusion dependent, thus forcing the classifiers to focus their decision on visible/non-occluded body parts. While

the paper introduces techniques for handling partial occlusions, it may still struggle with more severe occlusions where large portions of the pedestrian are obscured.

In order to detect partially occluded pedestrians, C. Zhou et al. [28] make use of the feature transformation technique by separating the features of pedestrians and non-pedestrian samples. In this approach, the authors use the feature space technique in which occluded pedestrians, non-occluded pedestrians, and background scenes are treated as three separate groups. The proposed technique requires carefully designed feature mappings, which can add computational complexity to the detection process. This complexity could result in slower processing times, particularly when applied to real-time systems.

A novel Mutual-Supervised Feature Modulation (MSFM) network has been proposed [21] which, instead of focusing only on the pronounced body features, compares all the features of the full body and visible body of the same pedestrian and calculates the similarity loss factor between the two pairs of the body features to help the detector learn the complete features of the pedestrian. The success of the Mutual-Supervised Feature Modulation approach is highly dependent on the quality of feature modulation during training.

In [22,29], the authors propose a pedestrian detector that detects a pedestrian sequentially in different stages. The first stage consists in identifying the locations of visible parts of the body that possess a high confidence score; the second stage involves balancing these visible parts by comparing them with the full body paradigm; and the final stage consists in analyzing these adjusted visible parts to decide the final decision about the target pedestrian. This approach is designed for handling static occlusions. The paper does not seem to address how the model might handle occlusions that evolve dynamically across different frames.

2.3. Attention Networks and Adversarial Learning Techniques

More often, pedestrian detection methods utilize body part detectors [14,30] to improve pedestrian detection performance, but these methods ignore the specific mechanism of combining the detection scores of the different body parts. A new approach has been proposed [31] with an enhanced study of the mechanism of merging the body part detectors by establishing two main processes: mining the body parts using a CNN-based mining model and combining the part detectors. Although the paper aims to address occlusion in pedestrian detection, the method may still struggle with severe occlusion cases where the pedestrian is highly obscured by other objects or pedestrians.

An attention-guided neural network model (AGNN) is proposed [32] for handling occlusion in pedestrian detection. A group of sub-images is generated with the help of a sliding window on the input image. From these sub-images, dominant features are extracted with the help of a convolutional neural network. In scenarios with highly occluded pedestrians or when the occlusion is non-uniform across the body, the attention mechanism might not be able to effectively focus on the most informative features.

Y. Pang et al. propose a novel network [33] comprising two blocks: a general pedestrian detection block, which detects the pedestrians by extracting the body features using full body annotations, and a mask-guided attention network, which exploits visible regions of the pedestrian and restrains the occluded parts of the body. In dense or highly cluttered environments, the mask-guided attention may result in the network incorrectly identifying background elements as pedestrians.

In [34], the authors utilize the fact that different channels of a CNN pedestrian detector are reactive to certain specific body parts, and the combination of such discrete body parts can be used to form a distinct occlusion pattern for each channel. The authors make use of a guided attention system to illustrate all such occlusion patterns in a single framework after performing optimization among them. The paper focuses on occluded pedestrian detection and re-identification but may struggle in crowded, multi-person scenarios.

S. Zhang et al. [35] propose a method that reduces the intra-class variance by extracting the complete features from the occluded regions so that the pedestrian features can be arranged between different occlusion patterns. The approach involves adversarial feature completion, which typically requires high-quality input data. If the input images are of low resolution, the adversarial completion might fail to recover accurate features, leading to degraded performance.

2.4. Pedestrian Detection in Highly Occluded or Dense Environments

Y. Zhang et al. [36] focused on enhancing pedestrian detection efficiency in highly overpopulated scenes. The model mutually detects the body and the head by dynamically learning the head–body ratio of the pedestrians during the training phase. The focus on pedestrian heads as a key feature for tracking in dense crowds is a strength in many situations. However, this method might be limited when non-head occlusions occur. Furthermore, in dense crowds, individuals may be very close to each other, making it difficult to distinguish between adjacent heads.

Z. Zhou et al. [37] proposed a framework to detect the position of pedestrians in parking lots in order to avoid collision accidents. This framework incorporates various information sources to predict pedestrian position and status information for vehicles for the safety of pedestrians in parking lots. The paper seems to focus heavily on parking spaces, which might not account for the diverse range of challenges faced in other occlusion-heavy environments.

In the available literature [38], a high degree of diversity is found in the interpretation of the severity level of occlusion because of certain factors, such as object size, distance of the object from the camera, lighting effects, and environmental conditions. To counter these multiple problems, Ref. [39] proposed an objective-based benchmark for a detailed characterization of the pedestrian detection techniques. The proposed benchmark covers the complete range of occlusion levels from 0% to 99%. The paper primarily focuses on creating an objective benchmark for occluded pedestrian detection and may not focus as much on the development or enhancement of the detection algorithms themselves.

M. Thu et al. [25] proposed a network that consists of two branches: the first branch is used to couple the outputs generated from the discriminative features of the pedestrians, and the second branch is used to integrate the discriminative features with the full body characteristics of the pedestrian. The performance of the TCNN architecture may depend heavily on the dataset used for training and evaluation. If the dataset is limited in terms of diversity, the model may struggle to generalize to other real-world scenarios.

2.5. Use of Vision Transformer (ViT) Model

Vision Transformers (ViTs) have lately developed as a competitive alternative to Convolutional Neural Networks (CNNs) in many image identification computer vision applications. The ViT model architecture was first introduced by Neil Houlsby and his team in a research paper [40] in 2021. The first step in processing an image with ViT is to split the image into fixed-size, non-overlapping patches. Each patch is then linearly embedded into a higher-dimensional space using a trainable linear projection. Position embeddings are then added to the patch embeddings to retain spatial information. The resulting combined sequence of patch embeddings and position embeddings is passed through a transformer encoder composed of several layers of self-attention and feedforward networks. After processing through the transformer encoder, the output can be used for various tasks. In [41], the authors propose using a Vision Transformer (ViT) model for modality fusion to address the problem of identifying complicated scenes by combining two distinct forms of remote sensing data: hyperspectral photography, which offers comprehensive spectral information,

and LiDAR data, which provides 3D spatial information. The accuracy of classification can be improved by combining these two modalities, particularly in complicated situations like dense forests or cities.

For small-sized traffic sign detection, the authors of [42] suggest a patch-wise training approach with the YOLOv3 network layer pruning to increase mean average precision and recall. The authors also propose an anchor box selection algorithm that uses bounding box dimension density to obtain an optimal anchor set for the dataset in order to decrease the log-average miss rate and false positives.

Sudhakar Tummala et al. in [43] make use of Vision Transformers (ViTs) for the classification of brain tumors based on Magnetic Resonance Imaging (MRI) scans. In contrast to conventional machine learning models, the study especially examines how combining many ViT models can increase the classification accuracy of brain tumor detection, outperforming traditional methods like CNNs and individual ViTs.

Recent development of multi-modal learning has created new research directions in important areas such as electricity price prediction [44], prediction for intelligent connected vehicles in smart cites [45], and mobile traffic prediction in consumer applications [46].

3. Materials and Methods

We designed a framework for occlusion handling in humans in which human samples from existing benchmark datasets are considered. Deformable patches from those samples are selected automatically through machine learning. We used deformation equations to correlate the deformable body parts of the human. These deformation equations are used to handle off-positions of the human body parts and decide how much more/less the body parts can deform. We used the proposed equations to detect deformable patches inside a human image, and score distribution analysis is performed to make the final call. We evaluated the performance of the model using a precision–recall curve, analyzing area under the curve (AUC), and the evaluation metrics of mean average precision 50 (mAP50).

3.1. Need for a Dedicated Occlusion Handling Model

Occluded objects offer a great challenge in the reliability of modern object detection systems. This necessitates the use of occlusion handling frameworks within a modern object detection system to obtain good system reliability. Most state-of-the-art methods use techniques like crops and mosaic augmentation and the You Only Look Once (YOLO) framework to handle occlusion.

In these methods, the object size is larger and similar in size to the image dimensions. However, objects that are relatively smaller than the image size offer a great challenge in predicting the object type. The problem could be handled by increasing the number of grid cells within the image, but this solution is accompanied by a complex computing analysis.

Accordingly, image tiling [47] has come out as a practical technique for the efficient detection of smaller objects as shown in Figure 3. In this technique, the input images are divided into overlapping tiles as shown in Figure 3a. The images are divided into smaller images with the help of overlapping tiles, where the tile sizes are chosen according to the image sizes used in the training system as shown in Figure 3b. Although this technique provides better results, there is a probability of detection of false positive images, especially when only a smaller part of an object is visible in a specific tile. This problem could be handled by using multiple bounding boxes in the final output image with the help of Non-Maximum Suppression (NMS) [48], but this does not guarantee a complete solution to the problem because of the estimation of bounding boxes of different sizes on the object of the final output image.



Figure 3. Image tiling. (**a**) Division of an input image into overlapping tiles. (**b**) Extraction of input tiles from the input image.

In order to meet these challenges, a customized occlusion handling framework is needed for the precise detection of pedestrians, which can be used in daily life applications. The use of such frameworks is also needed to cater to dynamic scenes where objects may be partially or occasionally occluded. This scenario requires highly advanced algorithms to precisely detect true objects from occluded segments. We propose a customized structure of an occlusion handling framework in practical systems that uses YOLO for object detection.

3.2. Proposed Occlusion Handling Network

We propose a novel occlusion handling framework that is designed to handle occlusion within objects that express high intra-class feature variations due to motion and other intrinsic characteristics. The main idea of our technique is to combine part localization with deformation-aware scoring via ML-learned regions. This is achieved by the identification and then classification of discriminative regions within the object using a discriminative deformable part model. For the identification of the discriminative regions, this model makes use of mathematical equations to isolate part models in correspondence to a central model, prohibiting any variation from this central location in order to preserve coherence between different discriminative parts.

The proposed framework consists of two main phases, i.e., identification phase and validation phase, as shown in Figure 4. The first block of our framework consists of the recognition of discriminative regions of the object. These regions are used to train a classifier on various parts of the object. This classifier, which is established as our discriminative deformable part model, incorporates mathematical equations to make sure that every single part model is strictly in coordination with a central model, thus strengthening the capability of the model to cater to intra-class feature variations due to motion and other inherent characteristics.

In order to detect the object inside an image, our framework chooses the most effective discriminative deep part model among a set of different deep mining models. Once the object is selected, the leftover discriminative part models are analyzed with the help of score distributions developed by the mathematical equations. These score distributions are then used to check the closeness of association of each part model with the detected object, thus enabling efficient recognition and classification.

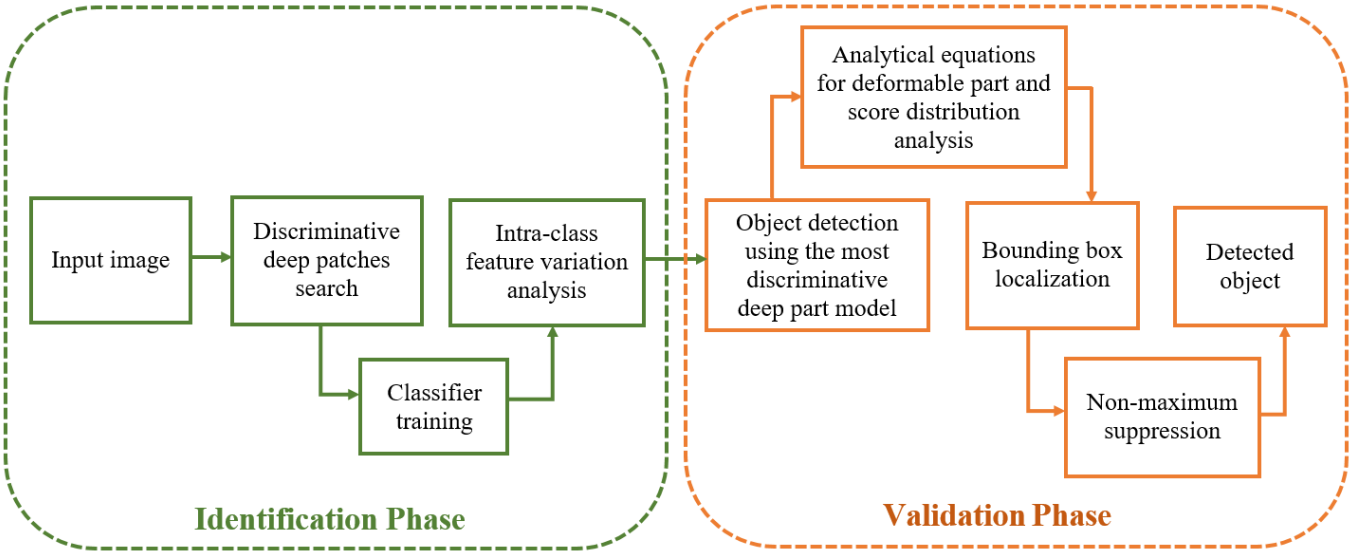

Figure 4. Proposed occlusion detection framework.

A voting technique is then used to finally check the validity of the detected object with the help of mathematical equations and score distribution analysis. If a larger proportion of the part model matches sharply with the detected object, the equation assigns it a relatively higher weight, and the object is categorized as a true positive; otherwise, it is categorized as a true negative. In this way, the use of our proposed discriminative deformable part model enables us to strengthen the accuracy and credibility of the object detection. The pseudocode for the presented algorithm is provided in Algorithm 1.

Algorithm 1 Algorithm to find the discriminative regions

Following is the description of variables:
D: Set of all positive square patches
A1, A2: Train set split
P: Original window
T_{acc}: Initial classification accuracy threshold
Δ_{inc}: Accuracy threshold increment
$Overlap_{max}$: Maximum allowed patch overlap

1: $A1 \leftarrow$ 80% training set, $A2 \leftarrow$ 20% validation set
2: $X \leftarrow \emptyset$ (Empty set of selected d-regions)
where $X \leftarrow$ Set of d-regions
3: $D = \{d \mid d \subset A1, \quad d \text{ is square}\}$
4: $\forall d \in D$, side(d) = $\frac{1}{2}$ x size(P)
5: $Overlap < Overlap_{max}$
6: $\forall d \in D$, Train classifier C_d
where Pos samples = patches from d
and Neg samples = patches from other regions
7: $A_{cc}(C_d) = \frac{\sum_{i \in A2} 1(C_d(F_i)=y_i)}{|A2|}$
where, 1 (.) is indicator function
$1(.) \leftarrow \begin{cases} 1 & \text{if } y_i \text{ is predicted correctly} \\ 0 & \text{otherwise} \end{cases}$
8: Patch selection: X = X U {d} if $A_{cc}(C_d) \geq T_{acc}$
$T_{acc} \leftarrow T_{acc} + \Delta_{inc}$
9: Otherwise: discard C_d
10: Iterative mining: $D_{new} = D \backslash X$
11: Termination criterion: if ($|A_{cc_{iter}} - A_{cc_{iter-1}}| \leq 3\%$ for 10 iterations)
or ($A_{cc} \geq 95\%$)
or ($|D_{new}| = 0$)

Apart from this, our proposed approach also takes into account those cases that are located on the borders of the tiling, causing remarkable refinement in detection accuracy on well-known benchmark datasets. Hence, by incorporating discriminative deep part models and using analytical equations to preserve dimensional coherence, our proposed framework demonstrates a greater strength in handling the detection of occluded objects within an image.

3.3. Discriminative Deep Part Models

The discriminative deep part model (DDPM) framework uses the technique of recognizing the discriminative regions of the object, thus strengthening the detection accuracy of the model and reducing the effects due to occlusion or deformity. This method involves specifically concentrating on discriminative regions during the testing phase by identifying the most informative regions within the object, thus enhancing the system's accuracy under formidable conditions where traditional detection techniques might become restricted.

To realize DDPM, for a 64 × 128 pedestrian (parent) image window, it is required to mine '*n*' regions from all possible locations as shown in Figure 5. Among them, only the square regions are assessed in order to avoid the complexities in the problem as shown in Figure 5a. During the training phase, the positive training dataset of the Microsoft Common Objects in Context (MSCOCO) pedestrian class is divided into two subgroups: A1 (80%) and A2 (20%). From the A1 subgroup, square regions are selected at random from the assigned location pool. The size of these square regions is chosen accordingly as half of the area of the parent image window in order to balance the useful information and occlusion contents. The extraction activity is controlled to preserve unique discriminative features in evenly distributed regions. The maximum overlap between the distributed regions is restricted to 30% so that the variance in the training dataset is secured.

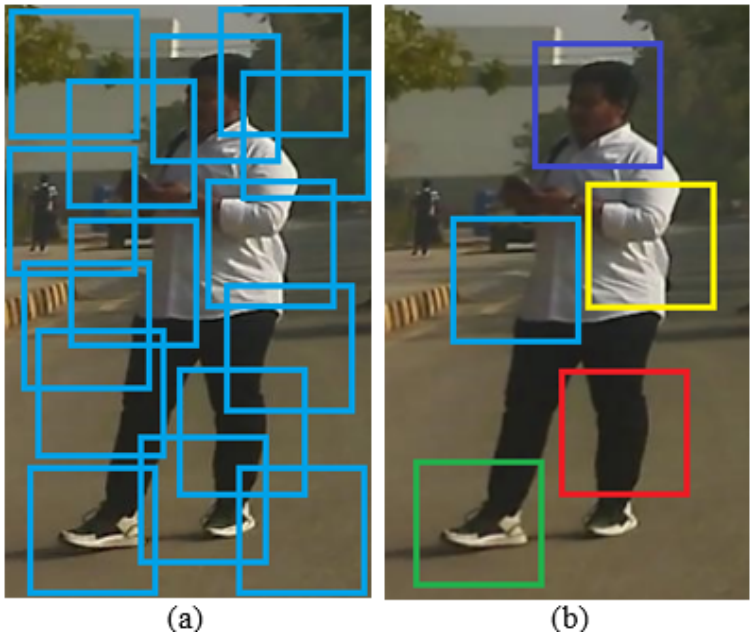

Figure 5. Extraction of the discriminative regions. (**a**) Candidate regions depicted inside the parent window. (**b**) Top discriminative regions identified inside the parent window.

After the activity of region selection, discriminative features are extracted from these regions with the help of the Darknet framework, which uses a one-vs-all strategy for the training phase. Figure 5b illustrates the recognition of top discriminative regions using this framework inside the MSCOCO pedestrian class, featuring the model's ability to handle occlusion in practical applications of object detection. In this strategy, the

samples of the target region group are taken as positive samples, while the remaining regions are taken as negative samples. The classifiers after training approach the A2 subgroup, and those classifiers with an accuracy greater than 50% are reviewed for extended improvement. If in this case none of the classifiers achieve these criteria, the whole operation is repeated by extracting regions anew and modifying the criteria in accordance with the performance metrics.

This process is repeated in a cyclic manner, until the classification accuracy becomes steady within predetermined deviation limits or surpasses the defined threshold criteria. The final output consists of a set of filtered classifiers labeled as discriminative deep part models. These classified models exhibit discriminative regions efficiently and contain distinguished properties related to the adjacent areas.

In order to practically implement this framework, immense computational methods were employed. It took around 200 h to undertake and extract the discriminative regions utilizing our CPU frequency of 2.5 GHz, PC RAM of 32 GB, and GPU GeForce RTX 3050, 8 GB, NVIDIA Corporation, Santa Clara, United States. For training, the batch size was set to 16. The initial learning rate was set to 0.001, and the momentum factor to 0.9. The total training duration comprised 8–15 epochs.

3.3.1. Deformable Equation and Score Distribution Analysis

In order to evaluate the decision weight W_k for discriminative deep part models, the following deformation equation is applied, which also serves to handle off positions of the deformable human body parts on the basis of which the model concludes how much more or less the body parts can deform.

$$W_k = \sum_{i=1}^{n} S_i e^{-(\frac{2d_i^2}{n^2})} \quad (1)$$

Here, S_i indicates the confidence score corresponding to the ith discriminative deep part model, d_i indicates the Euclidean distance of the ith model from the reference model, and n indicates the total number of discriminative deep part models. The decision weight given by Equation (1) determines whether to include a deep part model in the object detection mechanism or not, especially when handling deformable body parts.

We perform score distribution analysis using Equation (1) to control dimensional variations or deformations in discriminative deep part models corresponding to the reference model. The exponential term in Equation (1) serves as a factor that minimizes the effect of deep part models situated far away from the reference model, thus taking into account the deformations of the object. The factor of $\frac{2}{n^2}$ in the power of the exponent in Equation (1) manages the decline rate on the basis of Euclidean distance d_i so that the models situated close to the reference model have a higher impact on W_k compared to the far-away models.

This technique effectively controls the decision weight on the basis of dimensional variations in discriminative deep part models corresponding to the reference model, thus entitling an efficient handling of deformable objects and their off positions.

3.3.2. Bounding Box Estimation

Bounding boxes surrounding the objects are approximated by considering all the bounding boxes of the discriminative deep part models. The minimum and maximum values obtained from the part model detections are used to determine more efficiently the borders of detected objects. This technique customizes the Learning Cost Function (LCF) of YOLO frameworks, which performs a significant role in upgrading the accuracy of these approximations, resulting in exceptionally precise and valid estimates of pedestrian bounding boxes in occluded scenarios.

3.4. Our Proposed Dataset with Local Attire

Our proposed dataset focuses on the demanding situation of detecting pedestrians with Eastern attire, where humans normally wear clothes that cover the whole body, hindering the detector from recognizing the body parts. This dataset helps in improving the detector performance under various intra-class variations of the pedestrians. Bounding box distribution of pedestrians as recorded by the camera inside the car with local attire is shown in Figure 6.

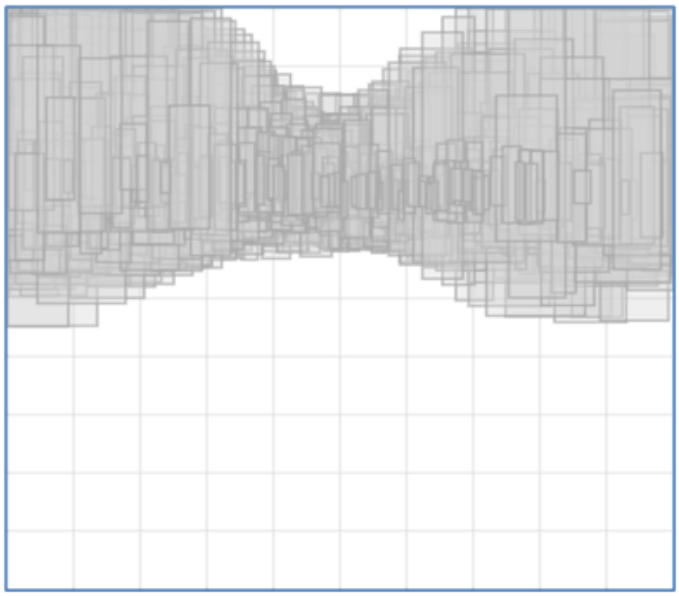

Figure 6. Bounding box distribution of pedestrians for the proposed dataset.

The proposed dataset has a total of 5609 images with a 1920 × 1080 image size, captured in an urban setting by mounting the camera on the front windshield of the car. A sample is considered occluded if 20% or more of it is obstructed; otherwise, it is classified as non-occluded. There were 2883 images in which pedestrians were captured. The data were annotated using the open-source Roboflow and all its standard tools. A total of 7091 annotations were performed, with an average of 2.459 annotations per pedestrian image. The images were processed at a speed of 12 frames per second, i.e., 83.33 ms were spent processing one image. The image distribution detail of the proposed dataset is presented in Table 1.

Table 1. Image distribution detail of the proposed dataset.

Description	Result
Total images in the dataset	5609
Occlusion threshold level	20%
Total images consisting of pedestrians	2883
Total number of annotations	7091
Annotations per pedestrian image	2.459
Image processing speed	12 fps
Image size	1920 × 1080

The dataset was split into three subsets, consisting of 4943 annotations in the training set (70% of the total annotations), 1470 annotations in the validation set (20% of the total annotations), and 678 annotations in the testing set (10% of the total annotations). This distribution is presented in Table 2.

Table 2. Annotation details of the proposed dataset in training, validation, and test sets.

Description	Result
Total number of annotations	7091
Annotations in the training set	4943
Annotations in the validation set	1470
Annotation in the test set	678

Our proposed dataset contains images of pedestrians covering a wide range of Eastern clothing intra-class variations. Figure 7 shows some sample images from our dataset of the pedestrians wearing common female Eastern attire, which almost covers the whole human body, making it very difficult for the detector to recognize the main discriminative body parts such as the head, legs, arms, and torso. Any dataset based on Western clothes certainly fails to detect in such an environment, drastically affecting its accuracy and performance.

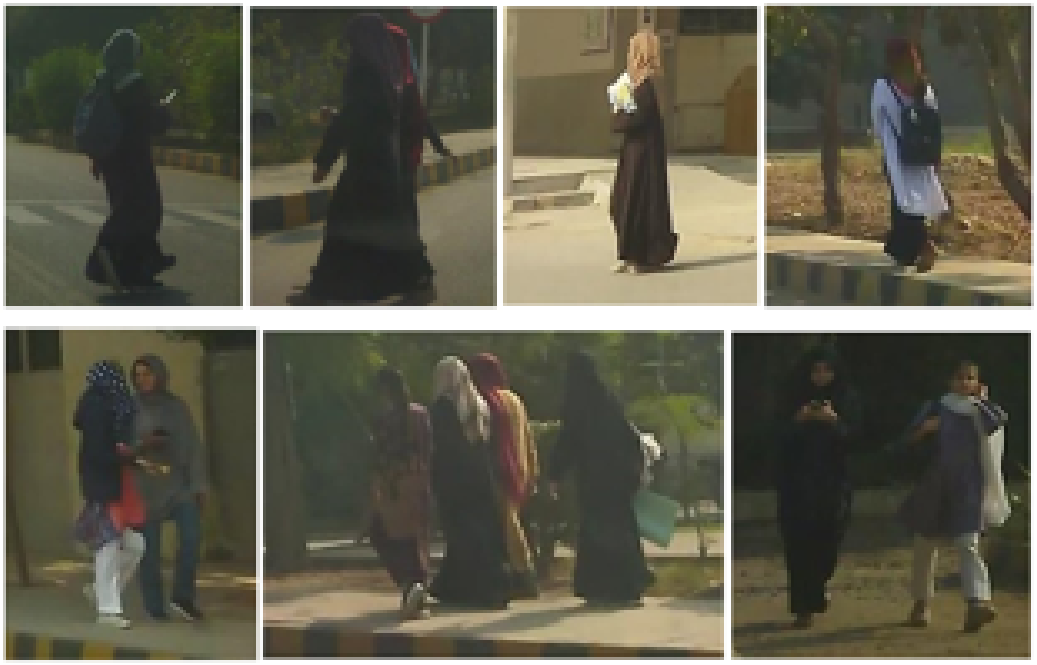

Figure 7. Sample images of the pedestrians from our proposed dataset with Eastern attire.

This dataset will play an important role in validating the performance of the detectors in the local settings and attire. This is achieved by using Transfer Learning, which can readily adapt to the local attire of the pedestrians.

Evaluation Metrics

The proposed dataset is evaluated on the Pascal VOC 2012 benchmark and the VisDrone 2019 detection test dataset using the precision–recall curve and evaluating area under the curve (*AUC*) using mAP50 evaluation metrics. The Mean Average Precision (mAP) is calculated by finding the average precision (*AP*) for each class and then averaging over several classes. If N is the total number of classes, the mAP is calculated using Equation (2).

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{2}$$

The mAP50 evaluation metric is calculated at an intersection-over-union (IoU) threshold of 0.5. The detections are matched by the ground truth until the overlap between both boxes is equal to or greater than 50% of the overlap area. The confidence threshold is

changed from lowest to highest value in the confidence score vector. For each change in the confidence score, true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are calculated. Once these values are calculated, they are used to evaluate the precision (PR) using Equation (3) and recall (RC) using Equation (4).

$$PR = \frac{TP}{TP + FP} \tag{3}$$

$$RC = \frac{TP}{TP + FN} \tag{4}$$

For each change in the confidence score threshold, one precision and one recall value are generated. All these values are then plotted on the precision–recall graph, and area under the curve (AUC) is calculated. If k is the number of points in the PR curve, the AUC is calculated using Equation (5).

$$AUC = \sum_{i=1}^{k-1} (RC_{i+1} - RC_i)(\frac{PR_{i+1} + PR_i}{2}) \tag{5}$$

The detector model that possesses the highest AUC is stronger than the other methods.

4. Experimental Results and Discussion

In this section, we show the detection performance of different algorithms using the datasets of Pascal VOC 2012 and VisDrone 2019 and compare our proposed network with the existing networks. We evaluate the proposed method using each dataset with the help of a precision–recall curve using mAP50 evaluation metrics. These metrics are used to estimate the accuracy of object detection models by calculating the precision and recall of the predicted bounding boxes compared to the ground truth.

4.1. Experiments with Pascal VOC 2012

The PASCAL Visual Object Class (VOC) 2012 dataset is a publicly available dataset. Pascal VOC 2012 is split into three subsets: 1464 images for training, 1449 images for validation, and a private testing set. Each image in this dataset has pixel-wise segmentation annotations, bounding box annotations, and object class annotations. Some sample images of the dataset are shown in Figure 8.



Figure 8. Sample images from Pascal VOC 2012 dataset.

On the Pascal VOC 2012 dataset, we compare our proposed method with You Only Look Once v11 (YOLOv11) [49], HF-YOLO [50], EFF-YOLO [51], and MSCD-YOLO [52] by calculating mAP50 values for the object category of person, and the detection results are summarized in Table 3, which shows that our proposed method achieves higher mAP

scores compared to the other techniques. Details on the accuracy advances in the person class on Pascal voc 2012 dataset can be found in [53].

Table 3. The mAP50 scores on PASCAL VOC 2012 testing dataset for object category of person.

Method	mAP Person
YOLOv11 [49]	61.5%
HF-YOLO [50]	81.60%
EFF-YOLO [51]	72.5%
MSCD-YOLO [52]	80.4%
DDPM (ours)	88.3%

4.2. Experiments with VISDRONE-DET2019

VisDrone-2019 is a publicly available benchmark dataset that consists of 288 video clips, formed by 261,908 frames and 10,209 static images, captured by drone-mounted cameras in different places at different heights as shown in Figure 9.



Figure 9. Sample images from VisDrone 2019 dataset.

The VisDrone-2019 dataset mainly focuses on humans and vehicles in our daily life and detects ten object categories of interest including *pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle*.

The dataset is divided into training, validation, and testing subsets (6471 for training, 548 for validation, 1580 for testing), which are collected from different locations but similar environments. One main feature of this dataset is that the object categories in the VisDrone data are highly unbalanced as shown in Table 4. For example, DPNet-ensemble produces 51.53% and 32.31% APs on the car and pedestrian classes while it only produces 18.41% and 12.86% APs on the awning-tricycle and bicycle classes.

Table 4. The AP scores on the VisDrone-DET2019 testing set of each object category [54]. The top three results are highlighted in red, green, and blue fonts for the person category.

Method	Ped.	Person	Bicycle	Car	Van	Truck	Tricycle	Awn.	Bus	Motor
DPNet-ensemble	32.31%	15.97%	12.86%	51.53%	39.80%	30.66%	30.66%	18.41%	38.45%	28.03%
RRNet	30.44%	14.85%	13.72%	51.43%	36.14%	35.22%	28.02%	19.00%	44.20%	25.85%
ACM-OD	30.75%	15.50%	10.26%	52.69%	38.93%	33.19%	26.96%	21.88%	41.39%	24.91%
BetterFPN	30.23%	16.45%	10.01%	51.45%	38.85%	31.57%	26.73%	17.79%	41.75%	24.83%
HRDet+	28.60%	14.58%	11.71%	49.46%	37.13%	35.20%	28.85%	21.93%	43.30%	23.55%

On the VisDrone 2019 dataset, we compare our proposed method with ACM-OD [54], DPNet-ensemble [54], and Better FPN [54], as these methods achieved the best three detection results for the person category, as shown highlighted in Table 4. DPNet-ensemble achieves a mAP of 15.97%, which is the highest among the listed methods in the table

after BetterFPN. ACM-OD achieves a mAP of 15.50%, which is slightly lower than DPNet-Ensemble and Better FPN. This result indicates that ACM-OD is performing well but is slightly behind the other two models in terms of precision across the detection tasks in the VisDrone-DET2019 dataset. BetterFPN achieves the highest mAP value of 16.45% in this table, indicating the best overall performance in terms of detection accuracy. Table 5 shows a comparison of our proposed method with these three methods. The detection results show that our proposed method achieved a 7.85% higher mAP score compared to the next best model of BetterFPN.

Table 5. The mAP50 scores on VisDrone19 detection testing dataset for object category of person.

Method	mAP Person
ACM-OD [54]	15.50%
DPNet-ensemble [54]	15.97%
Better FPN [54]	16.45%
DDPM (ours)	24.30%

4.3. Experiments with Our Proposed Dataset

To evaluate our proposed dataset, the images are collected from an urban environment. The most important thing is that it contains a large number of pedestrian samples with varying degree of occlusion and intra-class variations as shown in Figure 10.



Figure 10. Sample images from our proposed dataset, captured from an urban environment, showing a variety of intra-class variations in the pedestrians.

We first test the default data on our proposed dataset. These default data are based on Western attire. The precision–recall curve is plotted and is compared with that of YOLO v3 and YOLO v5. The comparison of the detection results is shown in Figure 11. Figure 11a shows a comparison between the AUC scores of our proposed model with those of YOLO v3 [55] and YOLO v5 [56], using the default data based on Western attire. The results show that our method achieves a higher performance score when trained on the dataset based on Western clothes.

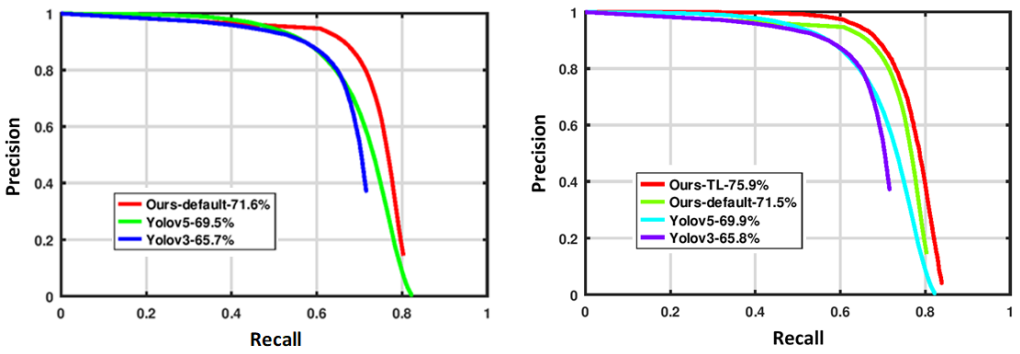

Figure 11. Detector evaluation on our proposed dataset by evaluating AUC. (**a**) Performance of YOLOv3, YOLOv5, and DDPM (ours) using default data. (**b**) Performance of YOLOv3, YOLOv5, and DDPM (ours) after transfer learning on local attire feature.

Transfer Learning

The data from the proposed dataset are Transfer Learned (TL) on the local attire features by employing the domain adaptation technique of fine-tuning in which pre-trained models can be fine-tuned on a small set of labeled target data to adapt to the target domain. The idea is to strengthen the knowledge that the model has learned during its training on the large dataset. Fine-tuning the model involves unfreezing some of the higher layers of the pre-trained model and training them with the new data. The initial layers of the pre-trained model are kept fixed, while the weights of the final layers are adjusted to adapt to local clothing variations.

The modified model, with frozen early layers and retrained final layers, is then trained on the new dataset during which the model learns to adjust the weights of the new layers while keeping the pre-trained weights in place. After fine-tuning, the model is evaluated on a test set for the new task. Figure 11b shows the comparison when the data from the proposed dataset have been Transfer Learned (TL) on the local attire features. The graph shows that our method again outperforms YOLO v3 and YOLO v5.

Figure 12 shows a comparison between sample images from the Pascal VOC benchmark dataset, the VisDrone19-Detection dataset, and our proposed dataset. Figure 12a shows images from the Pascal VOC dataset, Figure 12b from the VisDrone19 dataset, and Figure 12c from our proposed dataset.

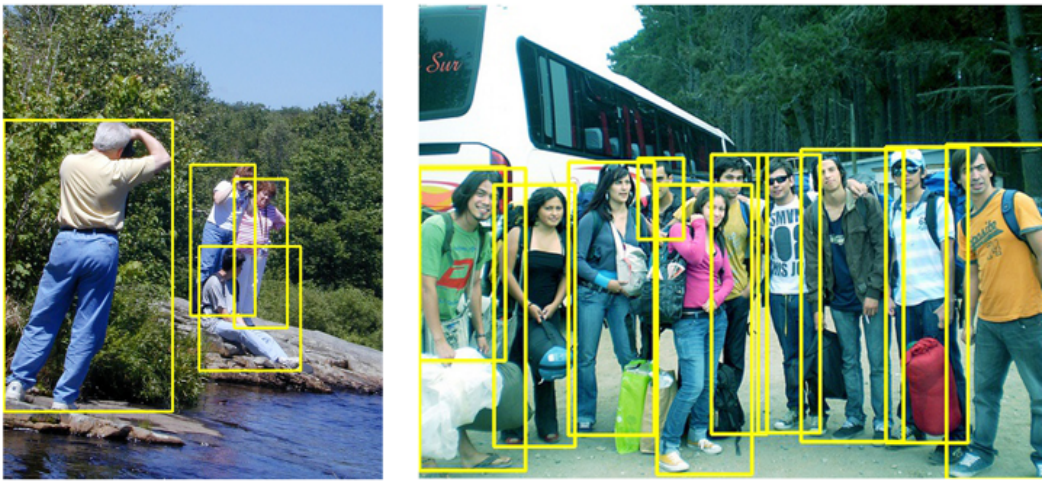

Figure 12. *Cont.*

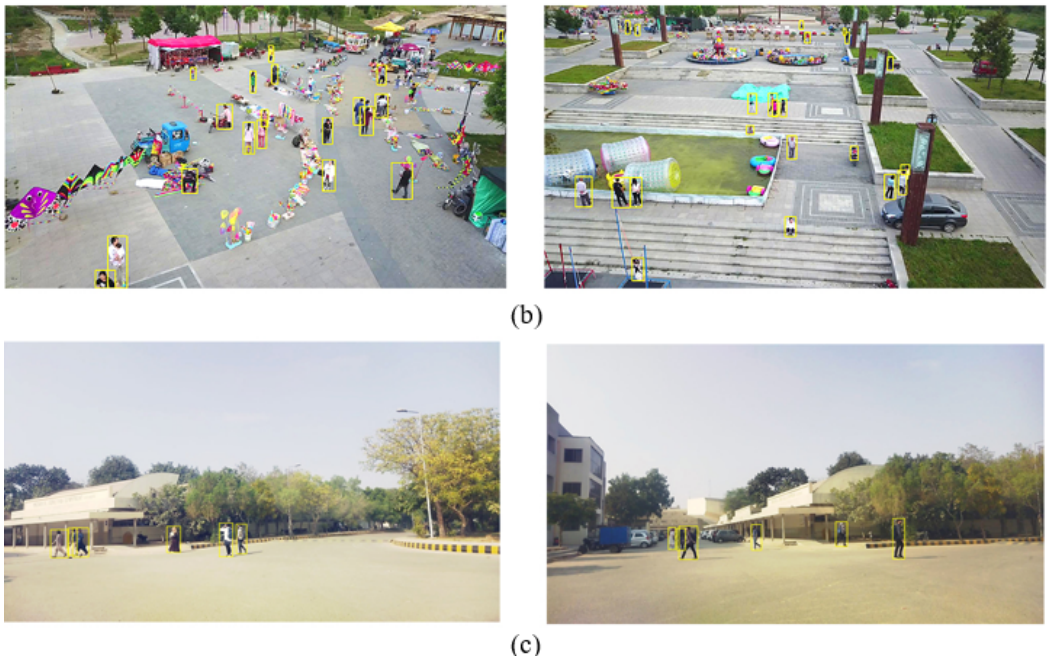

Figure 12. Comparison of Pascal VOC benchmark and VisDrone dataset sample images vs our method. (**a**) Sample images from Pascal VOC dataset. (**b**) Sample images from VisDrone dataset. (**c**) Sample images from our proposed dataset.

5. Conclusions

In this paper, we propose a new efficient occlusion handling technique for pedestrian detection that makes use of breaking a human body image into deformable parts through machine learning. The model automatically selects the discriminative deep patches to train the classifier on various parts of the body. We also proposed a new dataset recorded in urban Eastern culture environment to incorporate the intra-class variations in pedestrians which are very common in Eastern localizations. Using the default data, by finding area under the curve (AUC), our proposed algorithm with a detection accuracy of 71.6% outperforms YOLO v3 with 65.7% accuracy and YOLO v5 with 69.5% accuracy on the Pascal VOC benchmark dataset. After transferring learning, the algorithm on Eastern attire features, our proposed method achieves a detection accuracy of 75.9% compared to YOLO v3 with 65.8% accuracy and YOLO v5 with 69.9% accuracy. These experimental results indicate that the proposed approach is efficient and capable of outperforming other state-of-the-art methods for detecting occluded pedestrians. However, the main limitation of the study is that it does not account for variations in weather, lighting, or other environmental conditions that can significantly affect pedestrian detection performance. Extreme weather conditions, such as heavy rain, fog, or low light, might degrade the model's accuracy in outdoor settings. Future research should also consider the application of this study in the recent development of multi-modal learning, which is gaining more importance among researchers in areas such as electricity price prediction, prediction for intelligent connected vehicles in smart cites, and mobile traffic prediction in consumer applications.

Author Contributions: Conceptualization, S.S. and Y.R.; methodology, S.S.; validation, S.S., M.F.S. and Y.R.; formal analysis, S.S.; investigation, S.S. and Y.R.; data curation, S.S.; writing—original draft preparation, S.S.; writing—review and editing, S.S., M.F.S. and Y.R.; visualization, S.S. and M.F.S.; supervision, Y.R.; project administration, Y.R.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets used in this research are publicly accessible at the following links: Pascal VOC 2012 dataset at https://paperswithcode.com/dataset/pascal-voc-2012-test (accessed on 20 January 2025); VisDrone-DET2019 dataset at https://github.com/VisDrone/VisDrone-Dataset (accessed on 20 January 2025); YOLOv5 available online at https://github.com/ultralytics/yolov5 (accessed on 20 January 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1627–1645.
2. Rehman, Y.; Ahmed Khan, J.; Shin, H. Efficient coarser-to-fine holistic traffic sign detection for occlusion handling. *IET Image Process.* **2018**, *12*, 2229–2237.
3. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 743–761.
4. Dollár, P.; Appel, R.; Kienzle, W. Crosstalk cascades for frame-rate pedestrian detection. In Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 645–659.
5. Chen, N.; Li, M.; Yuan, H.; Su, X.; Li, Y. Survey of pedestrian detection with occlusion. *Complex Intell. Syst.* **2021**, *7*, 577–587.
6. Mohan, A.; Papageorgiou, C.; Poggio, T. Example-based object detection in images by components. *IEEE Trans. Pattern Anal. Machine Intell.* **2001**, *23*, 349–361.
7. Wu, B.; Nevatia, R. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Beijing, China, 17–21 October 2005; Volume 1, pp. 90–97.
8. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
9. Wang, X.; Han, T.X.; Yan, S. An HOG-LBP human detector with partial occlusion handling. In Proceedings of the 2009 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 32–39.
10. Enzweiler, M.; Eigenstetter, A.; Schiele, B.; Gavrila, D.M. Multi-cue pedestrian classification with partial occlusion handling. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 990–997.
11. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105
12. Mathias, M.; Benenson, R.; Timofte, R.; Van Gool, L. Handling occlusions with franken-classifiers. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 1505–1512.
13. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142–158.
14. Tian, Y.; Luo, P.; Wang, X.; Tang, X. Deep learning strong parts for pedestrian detection. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1904–1912.
15. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99.
17. Wang, X.; Xiao, T.; Jiang, Y.; Shao, S.; Sun, J.; Shen, C. Repulsion loss: Detecting pedestrians in a crowd. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7774–7783.
18. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 637–653.
19. Zhou, C.; Yuan, J. Multi-label learning of part detectors for occluded pedestrian detection. *Pattern Recognit.* **2019**, *86*, 99–111.
20. Sumi, A.; Santha, T. An Innovative Prediction Technique to Detect Pedestrian Crossing Using ARELM Technique. In *Human Behaviour Analysis Using Intelligent Systems*; Springer: Cham, Switzerland, 2020; pp. 119–140.
21. He, Y.; Zhu, C.; Yin, X.C. Mutual-supervised feature modulation network for occluded pedestrian detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 8453–8460.

22. Song, X.; Chen, B.; Li, P.; Wang, B.; Zhang, H. Prnet++: Learning towards generalized occluded pedestrian detection via progressive refinement network. *Neurocomputing* **2022**, *482*, 98–115.
23. Gilroy, S.; Glavin, M.; Jones, E.; Mullins, D. An objective method for pedestrian occlusion level classification. *Pattern Recognit. Letters.* **2022**, *164*, 96–103.
24. Huan, R.; Zhang, J.; Xie, C.; Liang, R.; Chen, P. MLFFCSP: A new anti-occlusion pedestrian detection network with multi-level feature fusion for small targets. *Multimed. Tools Appl.* **2023**, *82*, 29405–29430.
25. Thu, M.; Suvonvorn, N. TCNN Architecture for Partial Occlusion Handling in Pedestrian Classification. *Int. J. Pattern Recognit. Artif. Intell.* **2023**, *37*, 2350025.
26. Li, G.; Li, X.; Zhang, S.; Yang, J. Towards more reliable evaluation in pedestrian detection by rethinking "ignore regions". *Vis. Intell.* **2024**, *2*, 4.
27. Ouyang, W.; Zhou, H.; Li, H.; Li, Q.; Yan, J.; Wang, X. Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1874–1887.
28. Zhou, C.; Yang, M.; Yuan, J. Discriminative feature transformation for occluded pedestrian detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9557–9566.
29. Liu, T.; Luo, W.; Ma, L.; Huang, J.J.; Stathaki, T.; Dai, T. Coupled network for robust pedestrian detection with gated multi-layer feature extraction and deformable occlusion handling. *IEEE Trans. Image Process.* **2020**, *30*, 754–766.
30. Bourdev, L.; Yang, F.; Fergus, R. Deep poselets for human detection. *arXiv* **2014**, arXiv:1407.0717.
31. Cao, C.; Wang, Y.; Kato, J.; Zhang, G.; Mase, K. Solving occlusion problem in pedestrian detection by constructing discriminative part layers. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 91–99.
32. Zou, T.; Yang, S.; Zhang, Y.; Ye, M. Attention guided neural network models for occluded pedestrian detection. *Pattern Recognit. Lett.* **2020**, *131*, 91–97.
33. Pang, Y.; Xie, J.; Khan, M.H.; Anwer, R.M.; Khan, F.S.; Shao, L. Maskguided attention network for occluded pedestrian detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4967–4975.
34. Zhang, S.; Chen, D.; Yang, J.; Schiele, B. Guided attention in cnns for occluded pedestrian detection and re-identification. *Int. J. Comput. Vis.* **2021**, *129*, 1875–1892.
35. Zhang, S.; Ji, M.; Li, Y.; Yang, J. Imagine the Unseen: Occluded Pedestrian Detection via Adversarial Feature Completion. *arXiv* **2024**, arXiv:2405.01311.
36. Zhang, Y.; Chen, H.; Lai, Z.; Zhang, Z.; Yuan, D. Handling Heavy Occlusion in Dense Crowd Tracking by Focusing on the Heads. In Proceedings of the Australasian Joint Conference on Artificial Intelligence, Brisbane, QLD, Australia, 28 November–1 December 2023; Springer: Singapore, 2023; pp. 79–90.
37. Zhou, Z.; Yamada, S.; Watanabe, Y.; Takada, H. Tracking Pedestrians Under Occlusion in Parking Space. *Comput. Syst. Sci. Eng.* **2023**, *44*, 2109.
38. Gilroy, S.; Glavin, M.; Jones, E.; Mullins, D. Pedestrian occlusion level classification using keypoint detection and 2d body surface area estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3833–3839.
39. Gilroy, S.; Mullins, D.; Parsi, A.; Jones, E.; Glavin, M. Replacing the human driver: An objective benchmark for occluded pedestrian detection. *Biomim. Intell. Robot.* **2023**, *3*, 100115.
40. Dosovitskiy, A.; Beyer, L. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
41. Yang, B.; Wang, X.; Xing, Y. Modality fusion vision transformer for hyperspectral and LiDAR data collaborative classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 17052–17065.
42. Rehman, Y.; Amanullah, H.; Shirazi, M.A.; Kim, M.Y. Small traffic sign detection in big images: Searching needle in a hay. *IEEE Access* **2022**, *10*, 18667–18680.
43. Tummala, S.; Kadry, S.; Bukhari, S.A.C.; Rauf, H.T. Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *Curr. Oncol.* **2022**, *29*, 7498–7511. [CrossRef]
44. He, M.; Jiang, W.; Gu, W. TriChronoNet: Advancing electricity price prediction with Multi-module fusion. *Appl. Energy* **2024**, *371*, 123626.
45. Lu, Y.; Wang, W.; Bai, R.; Zhou, S.; Garg, L.; Bashir, A.K.; Jiang, W.; Hu, X. Hyper-relational interaction modeling in multi-modal trajectory prediction for intelligent connected vehicles in smart cites. *Inf. Fusion* **2025**, *114*, 102682.
46. Jiang, W.; Zhang, Y.; Han, H.; Huang, Z.; Li, Q.; Mu, J. Mobile traffic prediction in consumer applications: A multimodal deep learning approach. *IEEE Trans. Consum. Electron.* **2024**, *70*, 3425–3435.
47. Rehman, Y.; Amanullah, H.; Saqib Bhatti, D.M.; Toor, W.T.; Ahmad, M.; Mazzara, M. Detection of small size traffic signs using regressive anchor box selection and DBL layer tweaking in YOLOv3. *Appl. Sci.* **2021**, *11*, 11555. [CrossRef]

48. Akyon, F.C.; Altinuc, S.O.; Temizel, A. Slicing aided hyper inference and fine-tuning for small object detection. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 966–970.
49. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
50. Pan, L.; Diao, J.; Wang, Z.; Peng, S.; Zhao, C. Hf-yolo: Advanced pedestrian detection model with feature fusion and imbalance resolution. *Neural Process. Lett.* **2024**, *56*, 90.
51. Hu, M.; Zhang, Y.; Jiao, T.; Xue, H.; Wu, X.; Luo, J.; Han, S.; Lv, H. An Enhanced Feature-Fusion Network for Small-Scale Pedestrian Detection on Edge Devices. *Sensors* **2024**, *24*, 7308. [CrossRef] [PubMed]
52. Liu, Q.; Li, Z.; Zhang, L.; Deng, J. MSCD-YOLO: A Lightweight Dense Pedestrian Detection Model with Finer-Grained Feature Information Interaction. *Sensors* **2025**, *25*, 438. [CrossRef]
53. PASCAL VOC Dataset, Pedestrian Dataset. Available online: http://host.robots.ox.ac.uk:8080/leaderboard/displaylb_main.php?cls=person&challengeid=11&compid=3&submid=29793#KEY_refine_denseSSD (accessed on 18 January 2025).
54. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Lin, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
55. Redmon, J. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
56. YOLOv5. Available online: https://github.com/ultralytics/yolov5 (accessed on 18 January 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.