applied sciences

MDPI

Article

Occluded Pedestrian Detection Techniques by Deformable Attention-Guided Network (DAGN)

Han Xie, Wenqi Zheng and Hyunchul Shin *

Division of Electrical Engineering, Hanyang University, 55 Hanyangdeahak-ro, Ansan 15588, Gyeonggi-do, Korea; xiehan@hanyang.ac.kr (H.X.); zhengwenqi@hanyang.ac.kr (W.Z.)
* Correspondence: shin@hanyang.ac.kr; Tel.: +82-31-400-5176

Abstract: Although many deep-learning-based methods have achieved considerable detection performance for pedestrians with high visibility, their overall performances are still far from satisfactory, especially when heavily occluded instances are included. In this research, we have developed a novel pedestrian detector using a deformable attention-guided network (DAGN). Considering that pedestrians may be deformed with occlusions or under diverse poses, we have designed a deformable convolution with an attention module (DCAM) to sample from non-rigid locations, and obtained the attention feature map by aggregating global context information. Furthermore, the loss function was optimized to get accurate detection bounding boxes, by adopting complete-IoU loss for regression, and the distance IoU-NMS was used to refine the predicted boxes. Finally, a preprocessing technique based on tone mapping was applied to cope with the low visibility cases due to poor illumination. Extensive evaluations were conducted on three popular traffic datasets. Our method could decrease the log-average miss rate (MR^{-2}) by 12.44% and 7.8%, respectively, for the heavy occlusion and overall cases, when compared to the published state-of-the-art results of the Caltech pedestrian dataset. Of the CityPersons and EuroCity Persons datasets, our proposed method outperformed the current best results by about 5% in MR^{-2} for the heavy occlusion cases.

Keywords: pedestrian detection; feature extraction; computer vision; image processing

Citation: Xie, H.; Zheng, W.; Shin, H. Occluded Pedestrian Detection Techniques by Deformable Attention-Guided Network (DAGN). *Appl. Sci.* **2021**, *11*, 6025. https://doi.org/10.3390/app11136025

Academic Editor: Hugo Pedro Proença

Received: 14 May 2021
Accepted: 26 June 2021
Published: 29 June 2021

1. Introduction

Pedestrian detection is an essential computer vision problem that is widely utilized in many real-world applications, such as autonomous driving systems, robotics, and security monitoring systems. Inspired by deep-learning-based techniques of generic object detection, many research works [1–7] have achieved high detection accuracy for reasonable scale and non-occluded pedestrians. However, the detection performance is unsatisfactory for the difficult cases, such as crowd scenes, rare pose instances, and poor visibility cases influenced by time of the day or weather.

In traffic scenes, pedestrians are likely to be occluded by others or by roadside obstructions. Figure 1 shows several typical occluded cases in which the pedestrians are occluded by other pedestrians, trees, bushes, and cars parked on the roadside. In the Caltech pedestrian dataset [8], only 29% of pedestrians are never occluded, 53% are occluded in some frames, and 19% are occluded in all frames. One can notice that over 70% of pedestrians are occluded in at least one frame. In terms of the occlusion degree, 10% of pedestrians are "partially" occluded, and 35% are "heavily" occluded. Statistical analysis of the CityPersons dataset [9] indicates a similar situation in which fewer than 30% of pedestrians are not occluded. Since the occluded instances dominate the distribution, detecting pedestrians with occlusion is a critical issue that could considerably affect the overall detection performance. In this paper, we focus on improving the detection performance of occluded pedestrians in traffic scenes, mainly from three aspects: generating geometric transformation-invariant features for the deformed appearance of occluded

pedestrians, getting more accurate localization, and improving the test image quality when illumination is poor.

Figure 1. Examples of occluded pedestrians in traffic scenes. The green bounding boxes are the ground truth, the red bounding boxes denote the detections of our method.

Unlike rigid objects, the appearance and shape of pedestrians can be deformed under different poses and occlusions. Most deep neural networks adopt convolutional neural network (CNN) modules. However, the inherent attribute of a CNN unit is sampling the feature map from fixed locations. It is limited because different locations may correspond to pedestrians with different scales and poses. The deformable convolution network (DCN) [10] can adaptively decide the receptive field of the activation unit, and thus we embed it at high-level layers to encode more semantic information. Another important embedded feature is the self-attention module [11] which is designed in our work as an attention block. The attention block captures the dependencies between one position and others and re-weights the feature map with attention guidance. For image data, dependencies are captured by deep stacks of convolutional operations, in which the attention maps are formed within the local receptive fields. Benefiting from dependencies modeled by the attention mechanism, networks can flexibly adjust themselves to improve the representation ability. In our work, we capture the attention feature maps based on the shiftable receptive fields produced by the deformable convolution. The attention block can absorb more context dependency information and guide the network to pay more attention to pedestrian regions while suppressing background regions.

Another concern for pedestrian detection in a crowded environment is optimizing the prediction with accurate regression. A well-known evaluation metric is based on intersection over union (IoU), which considers the overlap of areas between two bounding boxes. However, most of the existing methods optimize the prediction by using the l_n-norm loss which computes the distance of two bounding boxes. There is a gap between prediction and evaluation. For instance, the l_2-norm loss of the detection box and the ground truth box are the same in Figure 2a,b, but the IoU values are different. A higher IoU value presents a better localization of predictions which is essential for the detection in crowded scenarios. The IoU-based regression loss, such as generalized-IoU [12] and distance-IoU [13] instead of l_n-norm loss, can produce more accurate localization of the detection bounding boxes. In addition, the IoU-based non-maximum suppression (NMS), which adaptively changes the threshold with the IoU of the neighboring detection bounding boxes, can refine the occluded detection boxes.

(**a**) l_2-norm = 12
IoU = 0.43

(**b**) l_2-norm = 12
IoU = 0.69

Figure 2. l_2-norm distances and IoU values of detection box (red) and ground truth box (green). (**a**) As an example, when l_2-norm is 12, the IoU of the two boxes is 0.43. (**b**) With the same l_2-norm, the IoU of the two boxes is 0.69. A higher IoU value indicates better detection localization.

Furthermore, illumination is a critical factor for detection performance. Since photographs are captured under diverse exposure, including overexposure or heavily shaded areas, the illumination frequently influences the visibility of pedestrians in test images. In the Caltech [8] and CityPersons [9] datasets, most of the images are collected in dry weather conditions, while in the EuroCity Persons dataset [14], many of the images were taken in rainy conditions with poor visibility. To cope with the challenging illumination cases, we capitalize a tone mapping technique as a preprocessing step. Our contributions are summarized as follows:

- First, we have designed a deformable convolution with attention module (DCAM) that generates the attention feature map corresponding to the deformable receptive field. The DCAM enables the network to adapt to diverse poses of pedestrians and occluded instances via deformable convolution. Furthermore, it can obtain attention features to capture effective contextual dependency information among different positions by a non-local (NL) attention block.
- Second, we have optimized the detection localization by using an improved loss function. The traditional smooth-L1 loss has been replaced with complete-IoU (CIoU) loss [13] for regression. The regression loss with CIoU, instead of the commonly used l_n-norm, can facilitate prediction with more accurate localization, as shown in Figure 2.
- Third, effective techniques for pedestrian detection in diverse traffic scenes have been explored in our work. The distance IoU-based (DIoU) NMS was adopted to refine the prediction boxes to improve the detection performance of occluded instances. A preprocessing with adaptive local tone mapping (ALTM) based on the Retinex [15] algorithm was implemented to enhance the detection accuracy under poor illuminance.
- Finally, experiments on three well-known traffic scene pedestrian benchmarks, Caltech [8], CityPersons [9], and EuroCity Persons (ECP) datasets [14], demonstrated that the proposed method leads to notable improvement in performance for the detection of heavily occluded pedestrians. Compared with the published best results, our proposed method achieved significant improvements of 12.44%, 5.3%, and 5.0%, respectively, in MR^{-2} of the heavily occluded sets of the Caltech [8], CityPersons [9], and ECP [14] datasets.

This paper is organized as follows: Section 2 reviews the existing closely related pedestrian detectors. Section 3 explains the details of our proposed method. Experimental results and ablation studies are presented in Section 4. Finally, conclusions are summarized in Section 5.

2. Related Works

2.1. Deep-Learning-Based Pedestrian Detection Methods

With the success of convolutional neural networks [16–19] in generic object detection, significant progress has been achieved in the pedestrian detection task. Most existing pedestrian detection methods are proposed using two-stages, based on a region-based

convolutional neural network (R-CNN) framework [18,20,21]. Two-stage pedestrian detectors generate a set of region proposals, and then classify the proposals into the pedestrian or the background classes and regress the coordinates in the second stage. For example, RPN+BF [4] uses the region-proposal network (RPN) to get candidate predictions which are then refined by the boosted forest (BF). For multi-scale pedestrian detection problems, MS-CNN [5] generates proposals by exploiting multi-scale feature maps. SA-Fast RCNN [22] contains two sub-networks for detecting pedestrians with a large- and a small-scale, respectively. DIF-RCNN [3] considers context information by integrating a deconvolution module, and enlarges the receptive field to enhance the detection performance of small-scale instances. One-stage pipelines have also been explored for pedestrian detection. ALF [6] is a lightweight pedestrian detector based on a single shot multi-box detector (SSD) architecture. It introduces an asymptotic localization fitting module that stacks multiple predictors to infer the anchor boxes step-by-step. The recent, state-of-the-art method, CSP [1] uses a single fully convolution network (FCN) with an anchor-free setting. It simplifies pedestrian detection as a center and scale prediction task.

2.2. Occluded Pedestrian Detection Methods

Regarding occluded pedestrian detection problems, a number of works adopt the part-based strategy that detects each part of a body and then fuses each prediction result to localize a partially occluded instance [23,24]. Some methods improve the performance of detecting occluded instances via developing an effective loss function. Rep-Loss [25] introduces a novel regression loss function to make the predicted candidate boxes less sensitive to the non-maximal suppress (NMS) threshold in crowded scenes. OR-CNN [26] takes advantage of the part-based strategy that integrates the body structure information with part occlusion-aware region-of-interest (RoI) pooling units, and designs a new aggregation loss function. Some other methods address the occlusion problem by optimizing the non-maximum suppression. For example, adaptive NMS [27] develops a density sub-network and applies a dynamic suppression threshold regarding the target density [28] and proposes an attribute map that encodes both the density and diversity information of crowd pedestrians, then designs an attribute-aware NMS algorithm to refine the detection results.

2.3. Attention- and Deformable-Convolution-Related Methods

Attention and deformable convolution have been proposed to enhance the representation capability of the network for object-detection and crowd-understanding fields. In [29], the deformable convolution is used in the context-feature-embedding module during the forward pass to obtain unevenly distributed context information, whereas the attention filter is applied during the backward pass. ADCrowdNet [30] proposes the attention map generator (AMG) to get the attention map, then uses the density map estimator (DME) with deformable convolution to generate the density map. In these methods, the attention and deformable convolution are developed in separate parts. On the contrary, the attention module is integrated into the DCAM in our approach to pay attention to the deformed appearance of occluded pedestrians.

In this work, a new deformable attention-guided pedestrian detector is proposed to achieve improved detection performance in occluded instances. Deformable convolution brings adaptive receptive field learning for pedestrians under different poses and occlusions. The non-local (NL) attention block is integrated to capture global context information. For classification and regression, an IoU-based loss function is used to optimize the model. Furthermore, instead of greedy NMS, we apply distance-IoU NMS (DIoU-NMS) [13], which generates the dynamic threshold with the IoU factor. This is to effectively suppress redundant boxes, which is useful in handling the detection of occluded instances. Additionally, adaptive local tone mapping [15] is implemented to further improve the detection performance by enhancing the visibility of objects with poor exposure.

3. Deformable Attention-Guided Network (DAGN)

The overall architecture of our proposed detector is illustrated in Figure 3. The baseline detector is the cascade R-CNN [31] with the structure of feature pyramid networks (FPN) [32]. For the feature extraction part, the deformable convolution with attention module (DCAM) is introduced with the backbone ResNet-50 [33]. The DCAM extracts rich context features in high-level layers with a deformable receptive field. In the detector head, the new optimized loss function replaces the conventional regression loss (l_n-norm) function with CIoU loss [13]. Then DIoU-NMS [13] is used to refine the bounding boxes in the crowded scenes. In order to overcome poor visibility problems in case of bad illuminance, adaptive local tone mapping (ALTM) based on Retinex [15] is adopted as a preprocessing step to further improve detection performance.

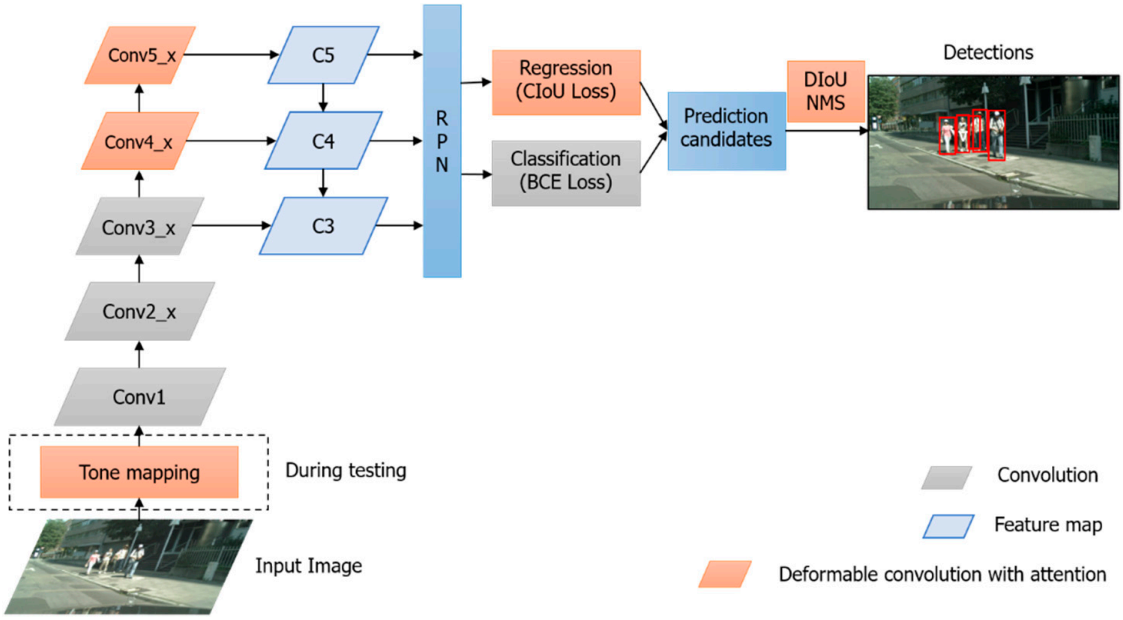


Figure 3. The architecture of our proposed deformable attention-guided network (DAGN) for pedestrian detection. The baseline detector is the cascade R-CNN [31] with the structure of feature pyramid networks (FPN) [32]. C3 to C5 denote the feature maps of the corresponding conv3 to conv5 stages of the ResNet50. The improved parts in our method are highlighted in red color.

3.1. Deformable Convolution with Attention Module (DCAM)

To augment the network capability of adapting to various appearances and poses of pedestrians, we designed the deformable convolution with attention module (DCAM), motivated by the deformable ConvNet v2 (DCNv2) [10] and simplified NL block [34]. The deformable convolution module enhances the capability of handling geometric transformation. To achieve the best trade-off between the performance and the efficiency, we just apply the DCAM at the conv4 and conv5 stages to minimize the computational cost. We designed the deformable convolution module based on DCNv2 [10]. Based on the preceding feature map, the offsets are learned by the deformable convolution layer to enable the non-rigid deformation of the sampling region. For each location p with sampling grid N, $x(p)$ is the input feature map and $y(p)$ is the output feature map. The deformable convolution module is defined with Equation (1):

$$y(p) = \sum_{p_n \in N} \omega_n \cdot x(p + p_n + \Delta p) \cdot mask \tag{1}$$

where ω_n denotes the weight for the n-th location, p_n enumerates the locations in sampling grid N, and Δp is the offset value of p_n. Thus the sampling is on the non-rigid locations

of $p_n + \Delta p$. A mask branch is designed by a sigmoid layer to decide whether to perceive signals from particular regions. The scalar *mask* lies in the range [0,1], Δp and *mask* are learnable values. Figure 4a shows the structure of the DCAM which is built in the conv5 stage of the ResNet50. For the conv4 stage, the kernel sizes and strides are kept the same, while the number of filters is halved. We replace the original convolution layer with the deformable convolution module to generate geometric transformation-invariant features. The deformable convolution module is denoted with the green dotted line in Figure 4a.

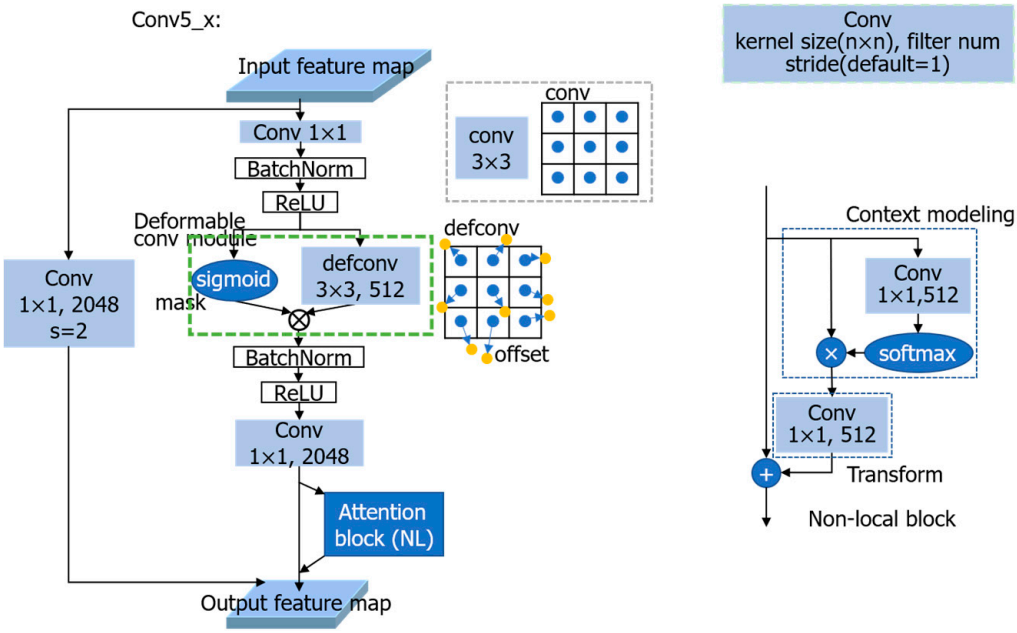


Figure 4. (**a**) The deformable convolution with attention module (DCAM) and (**b**) the non-local (NL) attention block built in the DCAM.

To further strengthen the global context information of the instances with different scales and deformed appearances, we incorporate the NL block with the deformable convolution module. Among the diverse self-attention mechanisms, such as squeeze-and-excitation (SE) [35], NL [36], global context (GC) [34], and convolutional block attention module (CBAM) [37], the NL block is superior for the pedestrian detection task. The NL block aims at capturing the dependency between two positions of a feature map in the spatial domain. These dependencies form the global context information to distinct the instance from the background. Other attention modules generate attention maps from the spatial domain, channel domain, and mixed domain. However, redundant information would weaken the discriminative ability of features, especially for the detection of small-scale instances.

Therefore, the NL block is incorporated in our model to capture the context dependency information. Due to the expensive computational cost of the original NL block, we adopt the simplified version of the NL block [34], as shown in Figure 4b. At first, a linear transform matrix (1×1 convolution) is introduced for global context modelling, followed by a SoftMax function to get attention weights. Then attention weights are applied to feature maps by using matrix multiplication to obtain the attention map. Next, we capture the global context dependency between the present position and all other positions using the above steps for each position. Then attention maps are transformed with a 1×1 convolution. Finally, the broadcast element-wise addition is employed to fuse the attention map with the feature for each position. By obtaining the global contextual dependency, the network can pay more attention to the features of the target position, thus discriminating the object from the background.

3.2. Target Optimization

3.2.1. Loss Function

Accurate localization is another important factor to improve pedestrian detection performance, especially in a crowded environment. To better optimize prediction localization, we apply the CIoU loss [13] to regress the predicted bounding box. The CIoU loss considers overlap area, central point distance, and aspect ratio, which are critical to measuring the similarity of the two boxes. It is defined as follows:

$$L_{CIoU} = 1 - IoU + R\left(B, B^{gt}\right) + \gamma v \tag{2}$$

where

$$IoU = \frac{\left|B \cap B^{gt}\right|}{\left|B \cup B^{gt}\right|} \tag{3}$$

$$R_{DIoU}\left(B, B^{gt}\right) = \frac{\rho^2\left(b, b^{gt}\right)}{c^2} \tag{4}$$

In the above, B and B^{gt}, respectively, denote the predicted box and target box, b and b^{gt} are the corresponding central points, $\rho(\cdot)$ is the Euclidean distance, $R_{DIoU}\left(B, B^{gt}\right)$ is the distance-IoU penalty term to minimize the normalized distance of the center points, c is the diagonal length of the smallest enclosing box covering B and B^{gt}, and γ is the trade-off parameter which is defined as $\frac{v}{(1-IoU)+v}$. We denote the width and height of the bounding boxes by w and h, respectively. The consistency of the aspect ratio of bounding boxes is computed by $v = \frac{4}{\pi^2}\left(tan^{-1}\frac{w^{gt}}{h^{gt}} - tan^{-1}\frac{w}{h}\right)^2$ as in [13].

For classification, we adopt the binary cross-entropy (BCE) loss as shown in Equations (5) and (6). The parameter p_i is the predicted probability and y_i is the ground truth label for the class.

$$L_{cls} = CE(\hat{p}_i) = -log(\hat{p}_i) \tag{5}$$

where

$$\hat{p}_i = \begin{cases} p_i & , if \ \ y_i = 1 \\ 1 - p_i & , otherwise. \end{cases} \tag{6}$$

To sum up, the overall objective function is derived as given in Equation (7), where λ is a trade-off coefficient, which is experimentally set to 5.

$$L_{total} = L_{CIoU} + \lambda L_{cls} \tag{7}$$

3.2.2. Non-Maximum Suppression for Prediction

Non-maximum suppression is used to suppress the redundant boxes and reject the false positive results. For NMS, DIoU-NMS [13] is adopted in our work. The DIoU-NMS penalizes the detection scores of neighbors with an adaptive threshold by the factor of the distance-IoU penalty term $R_{DIoU}(B_{max}, B_i)$, yielding better suppression for the occlusion cases. R_{DIoU} considers the distance between central points of a box B_i and the box with the highest score B_{max}. The DIoU-NMS is defined as follows:

$$s_i = \begin{cases} s_i, & IoU - R_{DIoU}(B_{max}, B_i) < \varepsilon, \\ 0, & IoU - R_{DIoU}(B_{max}, B_i) \geq \varepsilon, \end{cases} \tag{8}$$

where s_i is the classification score. The NMS threshold ε is experimentally set to 0.45.

3.3. Illumination Preprocessing for Testing

Uneven illumination of the test data is also a critical issue for pedestrian detection. Instances with low illumination often fail to be detected. To improve the illumination conditions of the testing images, we adopted a simple tone mapping method in the preprocessing procedure. In this work, we choose ALTM [15] for illumination preprocessing, which can improve the visibility of dark regions while keeping the detailed information

of bright regions. ALTM takes a small computational cost and can be easily integrated into the test pipeline. Other tone mapping techniques with similar properties can also be used. In ALTM [15], a global tone mapping is applied using Equation (9), according to the Weber–Fechner law [38]. $L_g(x,y)$ is the global adaptation output. N is the total number of pixels of the image. $L(x,y)$ is the input luminance. L_{max} denotes the maximum luminance value. $\overline{L}$ is the log-average luminance that is given as Equation (10). The guided filter is applied to the global adaptation to preserve the edge details, and is denoted by $H_g(x,y)$. The local adaptation $L_{out}(x,y)$ is computed by using Equation (11).

$$L_g(x,y) = \frac{\log\left(\frac{L(x,y)}{\overline{L}} + 1\right)}{\log\left(\frac{L_{max}}{\overline{L}} + 1\right)} \tag{9}$$

$$\overline{L} = \exp\left(\frac{1}{N}\sum_{x,y}\log(\delta + L(x,y))\right) \tag{10}$$

$$L_{out}(x,y) = \alpha(x,y)\log\left(\frac{L_g(x,y)}{H_g(x,y)} + \beta\right) \tag{11}$$

where δ is a small value to avoid the singularity of black pixels in the image, $\alpha(x,y) = 1 + \eta\frac{L_g(x,y)}{L_{gmax}}$ is the contrast enhancement factor, η is the contrast control parameter whose default value is 10, $\beta = \zeta\overline{L}_g$ is the adaptive nonlinearity offset relying on the log-average luminance of the global adaptation $\overline{L}_g$, and ζ is the nonlinearity control parameter (default $\zeta = 10$). Figure 5 presents two examples with the original test images (Figure 5a) and the preprocessed test images (Figure 5b) after applied ALTM. The detections of pedestrians missed in the original images can be detected after ALTM preprocessing, as shown in Figure 5c.

(**a**) (**b**) (**c**)

Figure 5. Detection comparisons before and after applying ALTM. (**a**) Original test images, (**b**) preprocessed images with tone mapping, and (**c**) detection of pedestrians. Red boxes denote the detections.

4. Experimental Results

In this section, we explain the details of the experimental setup and evaluation metrics. We evaluate the proposed method and make comparisons with the state-of-the-art methods of three popular pedestrian detection datasets, Caltech [8], CityPersons [9], and EuroCity Persons (ECP) [14]. The default evaluation settings of test subsets are shown in Table 1. The following sub-sections present the experiment details, including the ablation studies based on the Caltech heavy occlusion set.

Table 1. Benchmark datasets used for pedestrian detection.

Dataset	Case	Height of Pedestrian (pixels)	Occlusion Area	Visibility
Caltech [8]	Reasonable	>50	<35%	>0.65
	Heavy occlusion	>50	35–80%	0.2–0.65
	All	>20	-	>0.2
CityPersons [9]	Reasonable	>50	-	>0.65
	Heavy occlusion	>50	-	>0.9
	Partial occlusion	>50	-	0.65–0.9
	Bare	>50	-	<0.65
	Small	50–75	-	>0.65
	Medium	75–100	-	>0.65
	Large scale	>100	-	>0.65
EuroCity Persons [14]	Reasonable	>40	<40%	-
	Small	30–60	<40%	-
	Occluded	>40	40–80%	-
	All	>30	-	-

The range of occlusion area and visibility is from 0 to 1.

4.1. Experimental Setup and Evaluation Metrics

We implemented the proposed method based on the new parallel distributed deep learning framework PaddlePaddle [39] with version 2.0.0 and developed the code with PaddleDetection [40], an end-to-end object detection development kit based on PaddlePaddle. The experiments were performed in Python 3.7 and the Compute Unified Device Architecture (CUDA) with version 10.0. The ResNet-50 [33] pre-trained on the ImageNet [23] was used as our backbone. We used three parallel GTX Titan X GPUs during training and a single GTX Titan X GPU for testing.

For the evaluation of the pedestrian detection, we used the standard evaluation metric based on log average miss rate, over false positive per image (FPPI) range of 10^{-2} to 10^{0} (denoted as MR^{-2}) with the IoU threshold of 0.5. If the overlap ratio between the detected bounding box and the ground truth bounding box was less than 50%, the detected bounding box was determined as false positive. The lower value of the miss rate reflected the better detection performance.

4.2. Caltech Pedestrian Dataset

The Caltech pedestrian dataset is one of the most popular and large-scale datasets for the pedestrian detection task. It includes six train sets and five test sets in a sequence video format. In our experiment, we used the new annotations provided by [41] for training and testing. In total, 42,782 images are used for training and 4024 images for testing with the size of 480 $\times$ 640 pixels.

4.2.1. Training Configuration on Caltech Dataset

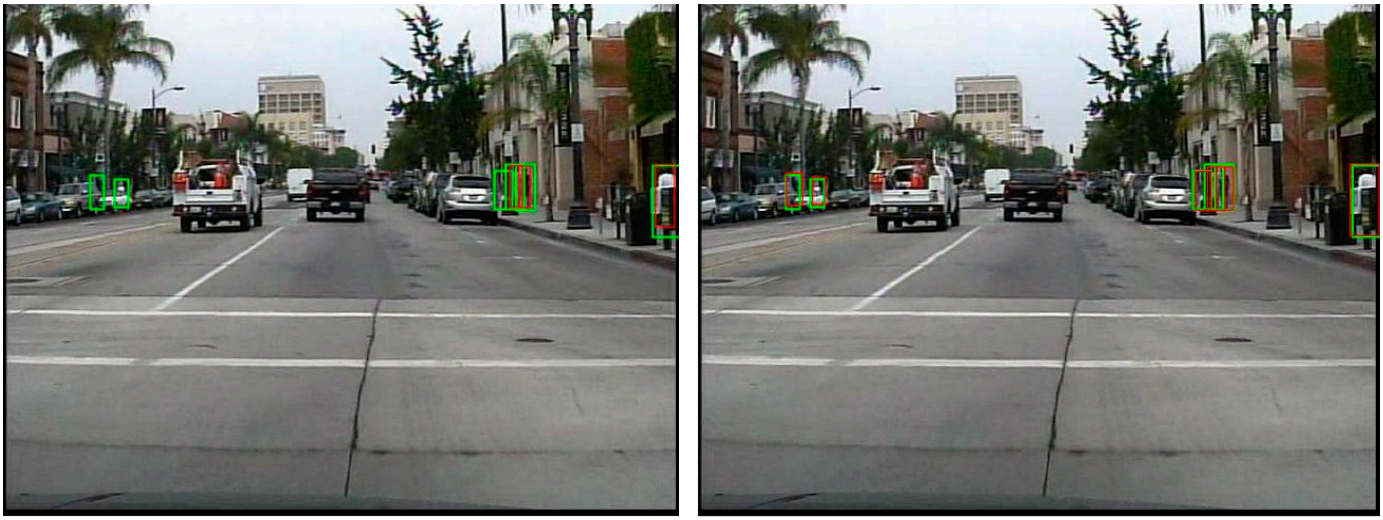
We used a multi-scale training strategy to detect different scales of pedestrians. During training, the images were resized, on the short side, into 11 scales (608, 640, 672, 704, 736, 768, 800, 864, 896, 928, and 960). In the region proposal network, the anchors were generated with multiple scales (16, 32, 64, 128, and 256), and the stride was set to 8. Because the bounding box aspect ratio distribution was 0.41 on average, we set the anchor scale as (0.41, 0.5, and 0.7). For the Caltech pedestrian dataset, the momentum was set to 0.9. The initial learning rate of 0.001 was used to optimize the model with a total of 70,000 iterations. After 55,000 iterations, the learning rate was reduced by a factor of 10, and after 62,000 iterations reduced again. The experiments were performed with a batch size of 2, both in training and evaluation. It took about 12 h to train the model.

4.2.2. Ablation Experiments on Caltech Pedestrian Dataset

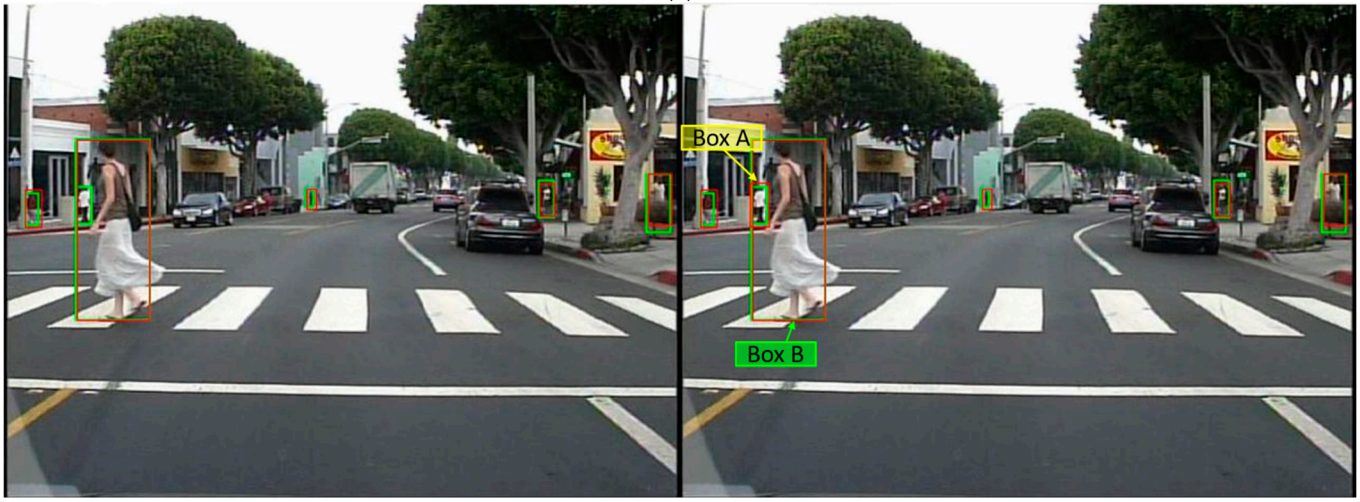
Ablation experiments were performed on the Caltech heavy occlusion test dataset to demonstrate the effect of each component added in our method. The following items were included: the deformable convolution with attention module (DCAM), the loss function with CIoU and DIoU-NMS, and the preprocessing of ALTM. The baseline method was cascade R-CNN+FPN. The detection performance was improved gradually when these modules were added one-by-one, as shown in Table 2. By applying the DCAM, an improvement of 8.51% in MR^{-2} was obtained compared to the baseline method. In Figure 6a, the occluded instances on the left side of the image with low contrast are difficult to distinguish from the background. On the right side of the image, the person occluded by the car failed to be detected by using the baseline method. However, these instances were all detected after using the DCAM, which demonstrates that our network can pay more attention to pedestrian regions even with occlusions, while reducing the background inference. Then, the designed loss function with CIoU optimized the prediction localization, which further improved the detection performance by 4.17% in MR^{-2}. By using the CIoU loss for regression, our detector could produce more accurate bounding boxes. DIoU-NMS helped to preserve the target boxes in crowded scenes. Figure 6b shows that box A with a high overlap area over box B was suppressed by the plain NMS, but box A remained with the DIoU-NMS. Thus, the predicted boxes of occluded pedestrians could be retrieved well. After adopting DIoU-NMS, the detection performance was improved by 5.48% in MR^{-2}. Moreover, the preprocessing using ALTM also contributed to an improvement of 3.92% in MR^{-2}, which provided a more robust detection performance under poor illumination.

Table 2. Influence of each component of our proposed method on the Caltech heavy occlusion subset.

Cascade R-CNN+FPN	DCAM (−8.51)	CIoU Loss (−4.17)	DIoU-NMS (−5.48)	ALTM (−3.92)	MR (%)
-					55.30
-	√				46.79
-	√	√			42.62
-	√	√	√		37.14
-	√	√	√	√	33.22



(**a**)



(**b**)

Figure 6. (**a**) Comparison of the detection performance based on the baseline method, cascade R-CNN+FPN. Left, without DCAM and right, with DCAM. (**b**) Detection examples with plain NMS (left) and DIoU-NMS (right). Green boxes, ground truth and red boxes, detection results.

4.2.3. Comparison with the State-of-the-Art Methods of Caltech Pedestrian Datasets

Table 3 shows the detection performances and runtime comparisons with the state-of-the-art methods, including MS-CNN [5], RPN+BF [4], SDS-RCNN [2], ALF [6], CSP [1], and Pedestron [42]. The progressive training pipeline proposed by Pedestron [42] is a good trick to improve the detection performance further. Since deep-learning-based methods depend heavily on the quantity and quality of data, progressively training the model from a large-scale and diverse training set to the relatively small dataset, which is closer to the target domain, can increase the representation and generalization ability of the model. ECP and CityPersons datasets are denser than the Caltech dataset, in terms of pedestrians per frame and are more diverse in scenarios with higher resolution. CityPersons contains 35,000 manually annotated persons with approximately 7 pedestrians per image on average. EuroCity Persons is nearly one order of magnitude larger with over 238,200 person instances manually labeled, and 9.5 pedestrians in average per image in crowded scenarios. From DAGN to DAGN++, we adopted a similar training strategy to that of Pedestron [42], i.e., pre-training the model from ECP, then fine-tuning from CityPersons.

Table 3. Comparisons with the state-of-the-art methods of the Caltech test set. R, reasonable; HO, heavy occlusion; and A, all. Scale ×1 presents original image scale was used for testing. Red color values denote the best results.

Method	Miss Rate			Hardware	Scale	Run Time (s/img)
	R	HO	A			
MS-CNN [5]	9.54	48.60	55.77	Titan GPU	×1	0.067
RPN+BF [4]	7.28	54.60	59.88	Tesla K40 GPU	×1.5	0.5
SDS-RCNN [2]	6.43	38.70	56.77	Titan × GPU	×1.5	0.21
ALF [6]	6.07	50.98	59.06	GTX1080 Ti GPU	×1	0.05
CSP [1]	4.54	45.81	56.94	GTX1080 Ti GPU	×1	0.058
Pedestron [42]	1.48	22.11	25.48	Nvidia tesla V100	×1	-
DAGN (ours)	6.03	33.22	46.83	Titan × GPU	×1	0.11
DAGN++ (ours)	1.84	9.67	17.68	Titan × GPU	×1	0.11

Our proposed DAGN achieved the lowest log-average miss rate of 33.22% on heavy occlusion cases without progressive training strategy, which outperforms the previous best result (38.70%) of SDS-RCNN [2]. For overall performance, our approach achieved 46.83% in MR^{-2} which exceeds the previous best result (55.77%) of CSP by a significant improvement of 8.94%. One can see that detecting occluded instances is essential for good overall detection performance. After applying the progressive pipeline, the DAGN++ also outperforms the state-of-the-art method, Pedestron, by 12.45% and 7.8% for heavy occlusion and overall performance, respectively, and achieves the competitive miss rate of 1.84%, which is 0.36% inferior to the best result (1.48%) of Pedestron for the reasonable test set.

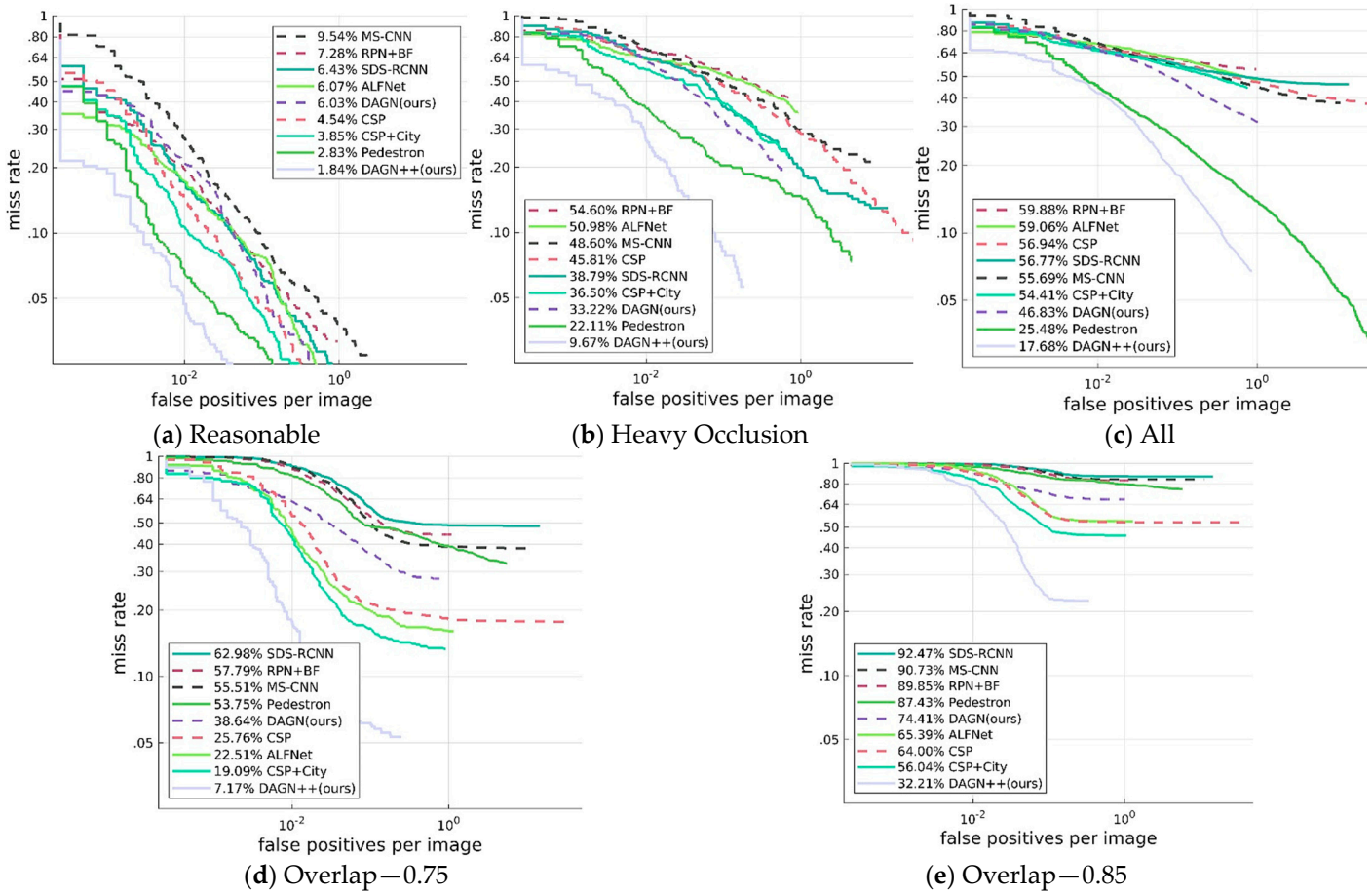
Figure 7 presents the detection miss rates versus FPPI with three simulations, namely "reasonable," "heavy occlusion," "all", "overlap—0.75", and "overlap—0.85" of the Caltech testing set. Compared to the standard IoU threshold of 0.5, the overlap ratio between the detected bounding box and the ground truth bounding box should be more than 75% and 85% to be considered as the true positive for "overlap – 0.75," and "overlap – 0.85" cases. Our proposed DAGN++ achieved the best results of 7.17% and 32.21% in MR^{-2} of the "overlap – 0.75" and "overlap – 0.85" cases, respectively. The results show that our method can produce more accurate localization. Figure 8 shows examples of the detection results of our method compared with the state-of-the-art method, Pedestron [42], of the Caltech pedestrian test set.

4.3. CityPersons Dataset

CityPersons [9] is a more diverse and challenging autonomous driving dataset. It collects images across 27 different cities with a high resolution of 2048 × 1024 pixels. Following [1,9,25], we conduct the experiment by using the official training set with 2975 images, and tested on the validation set with 500 images.

4.3.1. Training Configuration

During training, we also use a multi-scale training strategy. Considering the image size, batch size, and GPU memory, the images were resized, on the short side, into 11 scales (960, 992, 1024, 1056, 1088, 1120, 1152, 1184, 1216, 1248, and 1280). For the CityPersons dataset, the learning rate was decayed at 15,000 and 20,000 iterations with a total of 25,000 iterations. We used a batch size of 1 for training and a batch size of 2 for evaluation. It took about 6 h for training.



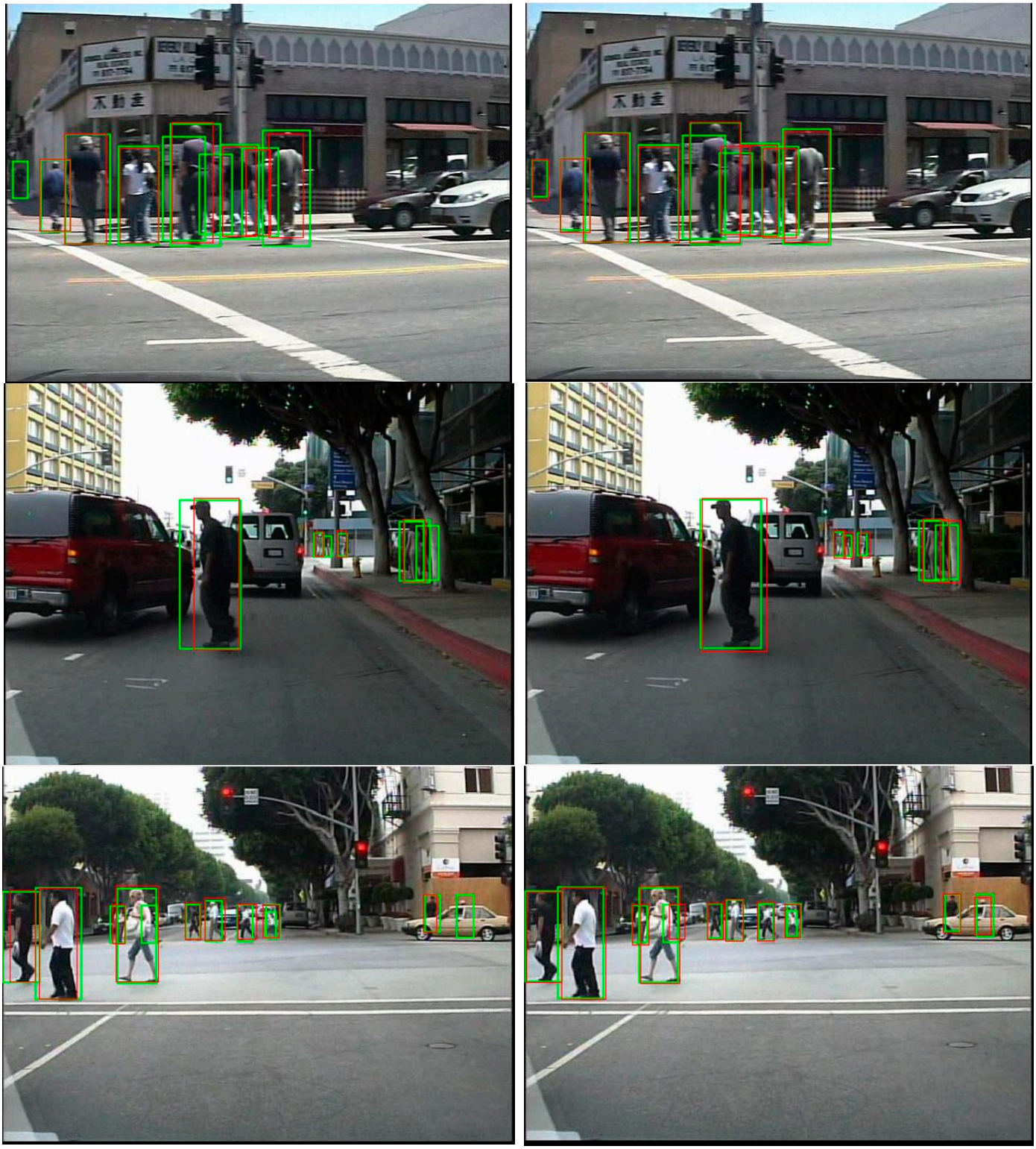
(**a**) Reasonable (**b**) Heavy Occlusion (**c**) All

(**d**) Overlap—0.75 (**e**) Overlap—0.85

Figure 7. Quantitative comparison results of the Caltech testing set with five subsets: (**a**) Reasonable. (**b**) Heavy occlusion. (**c**) All. (**d**) Overlap—0.75. (**e**) Overlap—0.85.

4.3.2. Comparison with the State-of-the-Art Method of CityPersons

Table 4 shows the comparisons with the state-of-the-art methods including Faster R-CNN [18], TLL [43], RepLoss [25], OR-CNN [26], ALF [6], CSP [1], Pedestron [42], and APD [28] on CityPersons. APD [28], OR-CNN [26], and RepLoss [25] are top pedestrian detection methods that focus on handling the occlusion problem. With the same ResNet-50 backbone, our proposed DAGN achieved 43.9% MR^{-2} which surpasses the second-best APD [28] (49.8%) by 5.9% in heavy occlusion cases. After being pretrained on the ECP, the performance of DAGN++ for heavy occlusion was 28.6% MR^{-2}, outperforming the previous best (33.9%) by a margin of 5.3% in MR^{-2}. For other cases, it also achieved competitive results. Figure 9 shows the visualization of the detection results of our method on the CityPersons validation set. One can find that several pedestrians missed by Pedestron [42] can be successfully detected by our proposed DAGN++.



(**a**) (**b**)

Figure 8. Example detection results of our proposed DAGN++ method compared with those of the Pedestron method of Caltech test set. (**a**) Pedestron [42] and (**b**) DAGN++ (ours). Green boxes, ground truth and red boxes, detection results. We visualized the detection boxes with the confidence score larger than 0.3.

Table 4. Comparisons with the state-of-the-art methods of the CityPersons validation set. R, reasonable; H, heavy occlusion; P, partial occlusion; B, bare; S, small scale; M, medium scale; and L, large scale. Red color values denote the best results.

Method	Backbone	R	H	P	B	S	M	L	Run Time (s/img)
Faster R-CNN [18]	VGG-16	15.4	-	-	-	25.6	7.2	7.9	-
TLL [43]	ResNet-50	15.5	53.6	17.2	10.0	-	-	-	-
RepLoss [25]	ResNet-50	13.2	56.9	16.8	7.6	-	-	-	-
OR-CNN [26]	VGG-16	12.8	55.7	15.3	6.7	-	-	-	-
ALF [6]	ResNet-50	12.0	51.9	11.4	8.4	19.0	5.7	6.6	0.27
CSP [1]	ResNet-50	11.0	49.3	10.4	7.3	16.0	3.7	6.5	0.33
APD [28]	ResNet-50	10.6	49.8	9.5	7.1	-	-	-	0.12
APD [28]	DLA-34	8.8	46.6	8.3	5.8	-	-	-	0.16
Pedestron [42]	HRNet	7.5	33.9	5.7	6.2	8.0	3.0	4.3	0.33
DAGN (ours)	ResNet-50	11.9	43.9	12.1	7.6	18.7	5.8	5.9	0.22
DAGN++ (ours)	ResNet-50	8.4	28.6	7.0	5.6	9.2	2.4	5.6	0.22

(**a**) (**b**)

Figure 9. Example detection results of our proposed DAGN++ method compared with those of the Pedestron method on the CityPersons validation set. (**a**) Pedestron [42] and (**b**) DAGN++ (ours). Green boxes, ground truth and red boxes, detection results. We visualized the detection boxes with the confidence score larger than 0.3.

4.4. EuroCity Persons (ECP) Dataset

The EuroCity Persons dataset [14] is a recently released large-scale and dense pedestrian dataset in urban traffic scenes, which is more diverse than Caltech and CityPersons. This dataset provides over 238,200 person instances with highly diverse and detailed annotations in over 47,300 images. The image size of the ECP dataset is 1920 × 1024 pixels. It records images including day- and night-time, and dry and rainy weather conditions across 31 different cities. For a fair comparison with other methods, we only used the 40,217 day-time images in the experiment. There were 23,229 images for training and 4225 images for validation. Following [14], we compared the results of the test set with 12,059 images.

4.4.1. Training Configuration

For the multi-scale training, we resized the images with the same setting as that of CityPersons. The momentum was set to 0.9. The initial learning rate of 0.001 was used to optimize the model with a total of 120,000 iterations. After 105,000 iterations, the learning rate was reduced by a factor of 10, and after 115,000 iterations reduced again. We used a batch size of 1 for training and a batch size of 2 for evaluation. It took about 28 h to train the model.

4.4.2. Comparison with State-of-the-Art ECP Dataset

We compared the detection results of our method on the ECP dataset with the existing reported results, as shown in Table 5. For the training and evaluation of the ECP dataset, we did not use other datasets for pre-training, and only evaluated the DAGN method. It took 0.22 s per image in testing. The proposed method outperformed other approaches for reasonable, occluded, and overall performance. In particular, for occluded cases, the DAGN achieved the miss rate of 26.3%, which clearly exceeds the second-best result of Cascade R-CNN [31] by 5%. The visualization of detection results is shown in Figure 10.

Table 5. Comparisons with the state-of-the-art methods in terms of log average miss rate on the ECP test set. Red color values denote the best results.

Method	Reasonable	Small	Occluded	All
SSD [16]	13.1	23.5	46.0	29.6
Faster R-CNN [18]	10.1	19.6	38.1	25.1
YOLOv3 [17]	9.7	18.6	40.1	24.2
Cascade R-CNN [31]	6.6	13.6	31.3	19.3
DAGN (ours)	5.9	14.2	26.3	17.5

4.5. Discussion

The experimental results and ablation study are described in Section 4. We evaluated the method of three popular datasets for autonomous driving to demonstrate the detection performance of occluded pedestrians in traffic scenes. Inspired by DCN [10], the deformable convolution technique was used to sample from non-rigid locations. We addressed the main bottleneck of the wide variation in the appearance of the person when occluded, by designing the deformable convolution network with an attention module. As a result, most of the occluded instances with the indistinguishable background could be successfully detected. Extensive experimental results showed that our combination of effective techniques, such as DCAM, CIoU loss, DIoU-NMS, and ALTM, can produce significantly improved results in "occluded" and "all" cases, when compared to those of several state-of-the-art methods.

In future work, we will consider applying our proposed pedestrian detector with lightweight network architectures to make the detection be real-time, while keeping the good detection performance. The real-time detector can be applied for the mobile and embedded vision applications, SSD [16] and YOLOV3 [17].

Figure 10. Example detection results of our proposed DAGN method on the ECP validation set. Green boxes, ground truth and red boxes, detection results. We visualized the detection boxes with the confidence score larger than 0.3.

5. Conclusions

In this research, a deformable attention-guided network (DAGN) for pedestrian detection was developed, in which deformable regions and attention features were used with global context information by proposing the DCAM. In order to optimize the prediction localization, an optimal loss function was designed, which combines BCE loss with CIoU loss for classification and regression. DIoU-NMS was followed to refine the prediction boxes, which further promotes the detection performance of occluded instances. Furthermore, the ALTM algorithm was applied as a preprocessing procedure to improve the detection performance under low illuminance conditions. Extensive evaluations demonstrated that the proposed DAGN achieves promising performance and outperforms other state-of-the-art methods, especially for heavily occluded pedestrians.

In future work, we plan to further decrease the computational cost and run time without sacrificing the performance. Network prune and distillation will be explored to make a lightweight, real-time pedestrian detector while keeping the detection accuracy.

Author Contributions: H.X. developed the idea, implemented the experiments, and wrote the manuscript. W.Z. prepared the Caltech dataset and participated ablation experiments. H.S. supervised the research and performed revisions and improvements. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program (10080619).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data in this study can be requested from the corresponding author.

Acknowledgments: This work was supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program (10080619).

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

References

1. Liu, W.; Liao, S.; Ren, W.; Hu, W.; Yu, Y. High-level semantic feature detection: A new perspective for pedestrian detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5182–5191.
2. Brazil, G.; Yin, X.; Liu, X. Illuminating Pedestrians via Simultaneous Detection and Segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4960–4969.
3. Xie, H.; Chen, Y.; Shin, H. Context-aware pedestrian detection especially for small-sized instances with Deconvolution Integrated Faster RCNN (DIF R-CNN). *Appl. Intell.* **2019**, *49*, 1200–1211. [CrossRef]
4. Zhang, L.; Lin, L.; Liang, X.; He, K. Is faster r-cnn doing well for pedestrian detection. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 443–457. [CrossRef]
5. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In *Proceedings of the Lecture Notes in Computer Science (Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2016; Volume 9908, pp. 354–370. [CrossRef]
6. Liu, W.; Liao, S.; Hu, W.; Liang, X.; Chen, X. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; Volume 11218, pp. 643–659. [CrossRef]
7. Xie, H.; Shin, H. Two-stream small-scale pedestrian detection network with feature aggregation for drone-view videos. *Multidimens. Syst. Signal Process.* **2021**, *32*, 897–913. [CrossRef]
8. Dollár, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 743–761. [CrossRef] [PubMed]
9. Zhang, S.; Benenson, R.; Schiele, B. CityPersons: A diverse dataset for pedestrian detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4457–4465.
10. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable convnets V2: More deformable, better results. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9300–9308.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
12. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
13. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7 February 2020; Volume 34, pp. 12993–13000. [CrossRef]
14. Braun, M.; Krebs, S.; Flohr, F.; Gavrila, D.M. EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1844–1861. [CrossRef] [PubMed]
15. Ahn, H.; Keum, B.; Kim, D.; Lee, H.S. Adaptive local tone mapping based on retinex for high dynamic range images. In Proceedings of the IEEE International Conference on Consumer Electronics, Las Vegas, NV, USA, 11–14 January 2013; pp. 153–156.
16. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European conference on computer vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37. [CrossRef]
17. Redmon, J.; Farhadi, A.; Ap, C. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
19. Ullah, I.; Manzo, M.; Shah, M.; Madden, M. Graph Convolutional Networks: Analysis, improvements and results. *arXiv* **2019**, arXiv:1912.09592.
20. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
21. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
22. Li, J.; Liang, X.; Shen, S.; Xu, T.; Feng, J.; Yan, S. Scale-Aware Fast R-CNN for Pedestrian Detection. *IEEE Trans. Multimed.* **2018**, *20*, 985–996. [CrossRef]
23. Wang, S.; Cheng, J.; Liu, H.; Tang, M. PCN: Part and context information for pedestrian detection with CNNs. In Proceedings of the British Machine Vision Conference, London, UK, 4–7 September 2017; pp. 1–13. [CrossRef]
24. 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.* **2018**, *40*, 1874–1887. [CrossRef] [PubMed]
25. Wang, X.; Xiao, T.; Jiang, Y.; Shao, S.; Sun, J.; Shen, C. Repulsion Loss: Detecting Pedestrians in a Crowd. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7774–7783.

26. 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, Munich, Germany, 8–14 September 2018; pp. 637–653.
27. Liu, S.; Huang, D.; Wang, Y. Adaptive NMS: Refining pedestrian detection in a crowd. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6452–6461.
28. Zhang, J.; Lin, L.; Zhu, J.; Li, Y.; Chen, Y.C.; Hu, Y.; Hoi, C.H.S. Attribute-aware Pedestrian Detection in a Crowd. *IEEE Trans. Multimed.* **2020**, *9210*, 1–13. [CrossRef]
29. Zhang, C.; Kim, J. Object detection with location-aware deformable convolution and backward attention filtering. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9444–9453. [CrossRef]
30. Liu, N.; Long, Y.; Zou, C.; Niu, Q.; Pan, L.; Wu, H. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3220–3229. [CrossRef]
31. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
32. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In Proceedings of the International Conference on Computer Vision Workshop, Seoul, Korea, 27–28 October 2019; pp. 1971–1980.
35. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef] [PubMed]
36. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
37. Woo1, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
38. Drago, F.; Myszkowski, K.; Annen, T.; Chiba, N. Adaptive Logarithmic Mapping for Displaying High Contrast Scenes. *Comput. Graph. Forum* **2003**, *22*, 419–426. [CrossRef]
39. Ma, Y.; Yu, D.; Wu, T.; Wang, H. PaddlePaddle: An Open-Source Deep Learning Platform from Industrial Practice. *Front. Data Comput.* **2019**, *1*, 105–115. [CrossRef]
40. PaddleDetection, v2.0.0-rc0. Available online: https://github.com/PaddlePaddle/PaddleDetection (accessed on 23 February 2021).
41. Zhang, S.; Benenson, R.; Omran, M.; Hosang, J.; Schiele, B. How far are we from solving pedestrian detection? In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1259–1267.
42. Hasan, I.; Liao, S.; Li, J.; Akram, S.U.; Shao, L. Pedestrian detection: The elephant in the room. *arXiv* **2020**, arXiv:2003.08799.
43. Song, T.; Sun, L.; Xie, D.; Sun, H.; Pu, S. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 536–551. [CrossRef]