electronics

MDPI

Article

Region Resolution Learning and Region Segmentation Learning with Overall and Body Part Perception for Pedestrian Detection

Yu Zhang, Hui Wang *, Yizhuo Liu and Mao Lu

The College of Computer Science and Engineering, Changchun University of Technology, Changchun 130012, China; zyzeroone@126.com (Y.Z.); yz_lisa@126.com (Y.L.); lumao@ccut.edu.cn (M.L.)
* Correspondence: email_wanghui@126.com

Abstract: Pedestrian detection is a great challenge, especially in complex and diverse occlusion environments. When a pedestrian is in an occlusion situation, the pedestrian visible part becomes incomplete, and the body bounding box contains part of the pedestrian, other objects and backgrounds. Based on this, we attempt different methods to help the detector learn more features of the pedestrian under different occlusion situations. First, we propose region resolution learning, which learns the pedestrian regions on the input image. Second, we propose fine-grained segmentation learning to learn the outline and shape of different parts of pedestrians. We propose an anchor-free approach that combines a pedestrian detector CSP, region Resolution learning and Segmentation learning (CSPRS). We help the detector to learn extra features. CSPRS provides another way to perceive pixels, outline and shapes in pedestrian areas. This detector includes region resolution learning, and segmentation learning helps the detector to locate pedestrians. By simply adding the region resolution learning branch and segmentation branch, CSPRS achieves good results. The experimental results show that both methods of learning pedestrian features improve performance. We evaluate our proposed detector CSPRS on the CityPersons benchmark, and the experiments show that CSPRS achieved 42.53% on the heavy subset on the CityPersons dataset.

Keywords: pedestrian detection; resolution learning; segmentation learning; computer vision; deep learning

Citation: Zhang, Y.; Wang, H.; Liu, Y.; Lu, M. Region Resolution Learning and Region Segmentation Learning with Overall and Body Part Perception for Pedestrian Detection. *Electronics* **2022**, *11*, 966. https://doi.org/10.3390/electronics11060966

Academic Editors: Jungong Han and Guiguang Ding

Received: 24 February 2022
Accepted: 19 March 2022
Published: 21 March 2022

1. Introduction

Pedestrian detection is the basis of other computer vision tasks, such as pedestrian re-identification, human pose estimation and other tasks. With the development of convolutional neural networks, pedestrian detection has gained improvements. With the development of deep learning and convolutional neural networks, there are two main methods for pedestrian detection, one is based on anchor-based approaches, such as Faster R-CNN [1], and the other is based on anchor-free approaches, such as CSP [2].

There are many improvements in pedestrian detection from the perspective of improving performance, such as fusing edge features to help pedestrian detection [3], merging segmentation features to assist pedestrian detection [3], utilizing part-based pedestrian detection with several parts of pedestrians [4,5], adopting visible features to improve performance [6,7], adopting semantic head to improve performance [8], combined with attention mechanism [9,10], cross-dataset training [11,12] and so on. Although there are many improvements in pedestrian detection, the performance still needs to be improved, especially in the heavy set. Therefore, it is important to improve the performance of pedestrian detection. The occlusion is complex and diverse, requiring highly representative tasks to guide the pedestrian detection task.

There are many improvements based on segmentation information, such as [3] and so on. The segmentation information contains the outlines and shapes of the pedestrian

in the pedestrian area. Therefore, the detector might benefit from these details regarding pedestrian features.

We do not fuse segmentation information to convolutional features. The detector has the ability to predict 1/4 the size of the segmentation map. The detector has the ability to distinguish between pedestrians and non-pedestrians. We adopt fine-grained segmentation annotations as the ground truth for the parts of each person, which includes edge information for pedestrians.

In [13], they proposed that the network should learn the input image more than once, the network should combine memory and input images to learn repeatedly at the different stages of the network. Instead, unlike them, we let the detector reconstruct the input image again. In the deep convolutional layers of the detector, the filters learn more abstract and blur features. For this problem, the filter contains details and structural features that are important for pedestrian localization.

We propose a new perspective that combines pedestrian detection with resolution learning. Ref. [14] proposed a method combining super-resolution learning and segmentation tasks. Inspired by their method, we propose to integrate region resolution learning into pedestrian detection. CSPRS unifies pedestrian detection and resolution learning.

The regional resolution pixel learning task assists the pedestrian detection task. Compared to pedestrian detection, this is a high feature representation task. This highly representative features guides and supervises pedestrian detection. We add the resolution learning branch to CSP [2]. The pedestrian detection task is further enhanced by the fine-grained structural representation. It is not involved in the inference stage and does not cost computation.

This resolution learning reconstructs details and structural features of images from the region pedestrian bounding boxes to learn pedestrian features. The region resolution learning branch is only sensitive to the region in pedestrian detection. Therefore, CSPRS has the ability to learn 1/4 the size of the original image input. CSPRS predicts the center, scale and 1/4 size image resolution pixel input at the same time. The detector not only learns the pedestrian position but also learns the resolution image input; therefore, the detector learns more features of each pedestrian.

The visible part of the pedestrian becomes incomplete in occluded environments, and the visibility of pedestrians is low in the case of severe occlusion; therefore, the detector has more difficulty predicting pedestrians. The regional resolution learning task focuses on regions within bounding boxes, which contain certain areas of pedestrians, some background and other objects under occlusion. Pedestrians are often in crowded scenes, and occlusion was divided into intra-class and inter-class occlusion. In the intra-class occlusion, especially in heavy intra-class occlusion, the detector usually detects pedestrians inaccurately.

Due to the diversity of objects within pedestrian bounding boxes, there is a weak ability to learn the diversity of other objects. Thus, the pedestrian detector should focus on the occluded pedestrians, even if the occluded pedestrians only account for a small percentage of the pedestrian bounding boxes. There are many kinds of occluded pedestrians, and we directly predict the pixel resolution pedestrian image to learn from the image itself. In this way, the detector learns from the input image to distinguish pedestrians from the background or other objects within the pedestrian bounding box area. By learning the regions of pedestrians, it might learn the silhouette and pedestrian silhouette when occluded. In order to make the detector focus on the pedestrian, we only predict the pixel region occupied by pedestrian images, and other parts are set zero.

Our main contributions include:

We introduce region resolution learning into the pedestrian detection field, which can make the detector more robust and contain more representation information. We combine the region resolution learning task with the pedestrian detection task, and the resolution learning task guides the pedestrian detection task. We propose a region resolution learning branch to keep 1/4 as the size of image resolution representation. In this way, the detector

maintains the high resolution of the image in the network fusing abstract features with high-resolution original features.

We propose a segmentation branch for pedestrian detection that predicts the segmentation of pedestrian parts on the pedestrian bounding box area. We perform different experiments to find a better style of adding resolution learning. After adding resolution learning, our proposed method CSPRS improves the performance.

2. Related Works

2.1. Anchor-Based

Faster RCNN [1] has been widely used in the pedestrian detection field, and there are many variations of Faster RCNN, such as adding an attention mechanism, adopting GIOU Loss, adopting new NMS loss and new feature fusion styles as well as other styles of modifiers.

There are many methods based on part-based approaches. In [4,5], the part-based detector first learns the parts individually and integrates the parts together. However, ref. [5] predicted the visible parts of five parts, and then fused the visible to predict full-body estimation. In [4], they combined a part-based approach with data enhancement; they randomly selected a part from five parts to add occlusions, thus, increasing the ability to handle occlusions.

There are some approaches that use visible part annotations to help locate full-body bounding boxes. Both [6,7] designed full bounding box prediction branch and visible part bounding box prediction branch and then fused these two bounding boxes, which were based Faster R-CNN. BCNet [15] also fused the full bounding box prediction branch and visible part bounding box prediction branch on the basis of CSP [2].

There are many methods that added attention mechanisms. On the basis of Faster R-CNN, ref. [9] added an attention module and a transform module. The attention module predicted the segmentation map, which consisted of 0 and 1. The value was set as 1 within the pedestrian bounding box and in others as 0. The transform module was adopted to handle occlusion, which separated the pedestrian and non-pedestrian.

There are many approaches to use segmentation and edge prior information. On the basis of Faster R-CNN, ref. [3] added a small branch, in this small branch, they attempted to define various channel features, including ICF channel, segmentation, edge, heat map, flow channel and depth features. Finally, only segmentation features helped to improve pedestrian detection performance. They merged this small branch and the backbone output features.

2.2. Anchor-Free

CSP [2] was first proposed in 2019 as an anchor-free detector; its architecture was simple and straightforward. It predicted the center and height of each pedestrian center. There are many improvements based on CSP [2], such as fusing the visible center and body center, modifying the feature fusion of the the four stage output, adding an attention module and so on.

BCNet [15] fuses the visible center and the body center, using the visible center to help refine the predicting body center. In PP-Net [16] adopts a new style of feature fusion. Recently, an attention mechanism was introduced into CSP [2]. In the [10], they proposed a spatial attention module and a channel attention module, which is attached after backbone output stage 4. The channel attention module acquired inter-dependencies between channels, and the spatial attention acquired long-range dependencies between pixels.

3. Proposed Method

3.1. Overall Architecture

The overall architecture is shown in the Figure 1, and the backbone adopts ResNeSt101 [17] to extract features. We designed a region resolution learning prediction branch to learn the

pixel image regions. CSPRS is based on CSP [2], which is an anchor-free detector. CSPRS architecture has two parts, the feature extraction and the detection head.

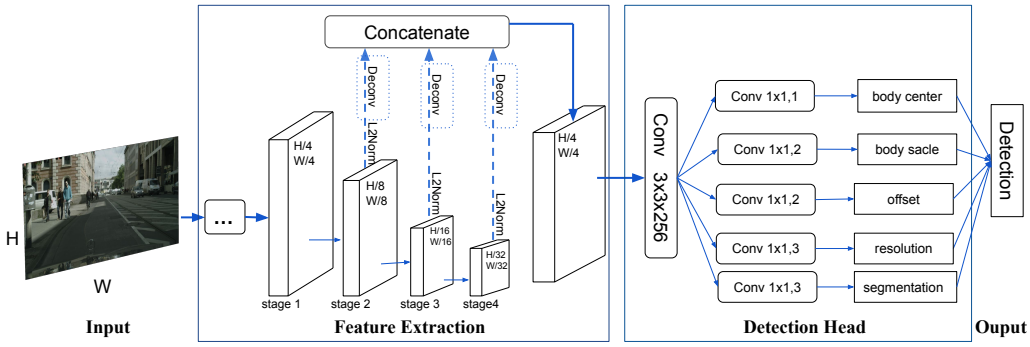

Figure 1. The architecture of CSPRS includes two parts: feature extraction and the detection head. The backbone of the detector is ResNeSt101 [17]. In the detection head, there are four output prediction branches, which are 1×1 convolution layers.

Feature extraction takes the image as input, and the output is a fused feature map at 1/4 the size of the image. The backbone extracts the input image features, we obtain the four stages output of the feature map, transpose these four feature maps, attach L2-normalization and concatenate these four feature maps. In the part of the detection head, the fusion feature attaches a convolutional layer with kernel size 3×3 and attaches four convolutional layers with kernel size 1×1 to predict the center, scale, offset, resolution region and segmentation. The offset output prediction is adopted to refine the position of the center, and the resolution output indicates region resolution learning prediction.

3.2. *Resolution Learning*

We live in a three-dimensional world: a point contains less information than a line, and a line contains less information than a surface; thus, it is important to find a high representation task to assist pedestrian detection. In this way, the high representation task contains more detailed features, and it is easy to infer low representation tasks.

Ref. [14] proposed a segmentation method on the basis of the results of previous studies. This method adopts super-resolution learning to improve segmentation learning. Considering why they adopted this method, the super-resolution learning task is a high task representation compared to the segmentation task. However, what task is highly representative compared to pedestrian detection? Considering this question, we predict not only the center and its height but also the image pixels of the pedestrian. In this way, the high representation task is helpful for low representation pedestrian detection.

Resolution learning also helps the network to better learn the attribute representation of a pedestrian. On the deep convolutional layers, the detector learns high-level features, which are abstract and blurred, and thus it might require the resolution pixel input feature to obtain an overall and detailed look.

Different from super-resolution learning, we only learn the pedestrian region resolution feature. In this way, resolution learning might learn the size and appearance of pedestrians, and the detector only focuses on the region of pedestrians rather than the background and other objects. As there are different objects and backgrounds that do not assist in pedestrian detection. In the output prediction, given an input, the detector outputs 1/4 the size of the input pedestrian region.

Our region resolution learning output prediction learns the region of pedestrian itself, the resolution prediction is 1/4 the size of the input size, The channel of resolution is 1 or 3, and the region is the pedestrian region or the total input image. The different strategies can be seen in Table 2. In total, we add a prediction output branch to the CSP [2], and the prediction learns the pixel image. We suggest that resolution learning is high representation compared with pedestrian centers and scale learning. Detectors simultaneously predict the

region resolution learning, center, scale and offset. We adopt a convolution layer to predict region resolution learning.

3.3. Segmentation Learning

Segmentation information contains outline and shape information, and fine-grained segmentation annotation is shown in Figure 2, which contains the different segmentation information for different body parts. Thus, the detector has the ability to distinguish different body parts of a person. The detector only learns the pedestrian region of the segmentation map, and thus it only focuses on the pedestrian area.

Figure 2. Examples of fine-grained segmentation. This picture comes from Cityscapes Panoptic Parts [18].

In [3], the method fuses extra features into convolutional features and the extra feature as prior information. We let the detector learn segmentation information. In this way, the detector can learn where the legs, head, arms, etc. are. The detector has the ability to obtain many shapes of pedestrian parts.

3.4. Training

3.4.1. Loss Function

The loss function has four parts: one is the pedestrian center, the second is the pedestrian size at the center position, the third is the offset, and the last is the resolution learning.

3.4.2. Ground Truth

For the label of the body center, we adopt 2D Gaussian, which can be seen in Equation (1). In this heat map, the center of the pedestrian is defined as positive, while the other points are defined as negative. At the center of the pedestrian, the value is set to 1. In other regions within the pedestrian bounding box, the value is set to the 2D Gaussian value. For the label of scale, the scale consists of height and width. We predict the height and width of each pedestrian at the center point, take the logarithm of height and width as the scale label. In the ground truth of the width map, within 2 pixels around the center point, the value is set as the logarithm of the width value.

The height map is similar to the width map. For the label of resolution learning, we adopt bilinear interpolation to resize the input image to 1/4 the size of an input image. The 1/4 size of an input image set as the ground truth, we only adopt the R channel of the input image, which has three RGB channels. For the label of segmentation learning, we adopt bilinear interpolation to resize the fine-grained segmentation map to 1/4 the size of the segmentation map.

3.4.3. Center Loss

The center loss adopts 2D Gaussian and Focal loss [19] to position the pedestrian in the input image—the 2D Gaussian function as seen in Equation (1). The Focal loss [19] is seen in Equation (2).

$$\begin{aligned} M_{ij} &= \max_{k=1,2,\ldots K} G(i,j;x_k,y_k,\sigma_{w_k},\sigma_{h_k}) \\ G(i,j;,x,y,\sigma_w,\sigma_h) &= e^{-(\frac{(i-x)^2}{2\sigma_w^2}+\frac{(j-y)^2}{2\sigma_h^2})} \end{aligned} \tag{1}$$

where K is the total number of pedestrians in each image, $(x_k, y_k, \sigma_{w_k}, \sigma_{h_k})$ are the position label of pedestrian. At the 2D Gaussian overlap location, we apply the element-wise maximum [2] value.

$$L_{center} = -\frac{1}{K}\sum_{i=1}^{\frac{W}{r}}\sum_{j=1}^{\frac{H}{r}}\alpha_{ij}(1-\hat{p}_{ij})^{\gamma}log(\hat{p}_{ij}) \tag{2}$$

where

$$\begin{aligned} \hat{p}_{ij} &= \begin{cases} p_{ij} & \text{if } y_{ij}=1 \\ 1-p_{ij} & \text{otherwise} \end{cases} \\ \alpha_{ij} &= \begin{cases} 1 & \text{if } y_{ij}=1 \\ (1-M_{ij})^{\beta} & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

In the above, (i,j) represent the location, y_{ij} indicate center or not. p_{ij} represents the prediction confidence score at location. The size of predicting map is $(\frac{H}{r}, \frac{W}{r})$. The σ is set as 2, and β is set as 4, which are suggested in [20].

3.4.4. Scale Loss

We define the scale prediction as a regression task via $L1$ loss [21],

$$L_{height} = \frac{1}{K}\sum_{k=1}^{K} L1(s_k, t_k) \tag{4}$$

where s_k and t_k indicate the height output prediction and ground truth height value of each positive pedestrian. We only focus on the height or width on the center location region, other locations are not considered. The total loss function contains L_{height} and L_{width}.

$$L_{scale} = L_{height} + L_{width} \tag{5}$$

3.4.5. Offset Loss

We define offset prediction as a regression task via smooth $L1$ loss [21].

$$L_{offset} = \frac{1}{K}\sum_{k=1}^{K} \text{Smooth}L1(o_k, \hat{o}_k) \tag{6}$$

where o_k and $\hat{o}_k$ represent the ground truth offset and prediction offset, and the offset loss is used to refine the center coordinates of each pedestrian.

3.4.6. Resolution Loss

We define this resolution learning task as a regression task. The size of the resolution prediction is 1/4 the size of the input, and we use Mean Squared Error (MSE) loss in Equation (7) to refine the resolution learning branch.

$$L_r = \frac{1}{A}\text{MSE}Loss(y_t, y_p) \tag{7}$$

where A represents the total pixels within the pedestrian bounding box, which indicates the body box instead of the visible part box. The y_t and y_p represent the pixel values of the

image within the pedestrian bounding box and the prediction of resolution learning. In this way, the detector only focuses on full-body bounding boxes containing the pedestrian and a few backgrounds when in the occlusion situation.

3.4.7. Segmentation Loss

We define the segmentation loss task as a regression task. The size of the segmentation branch is 1/4 of the input size. We use the Mean Squared Error (MSE) loss in Equation (8) to refine the segmentation learning branch.

$$L_s = \frac{1}{A}\text{MSE}Loss(y_t, y_p) \tag{8}$$

where A represents the total segmentation pixels within the pedestrian bounding box. y_t and y_p represent the segmentation map of pedestrians within the pedestrian bounding box and the prediction of the segmentation map.

3.4.8. Total Function

The total loss contains several parts, the loss of center, scale, offset, resolution learning and segmentation learning.

$$L = \lambda_c L_{center} + \lambda_s L_{scale} + \lambda_o L_{offset} + \lambda_r L_r + \lambda_s L_s \tag{9}$$

where λ_c is set as 0.01, λ_s is set as 0.05, λ_o is set as 0.1, λ_s is set as 0.0001 and λ_r is set as 0.0001.

3.5. Inference

During inference, resolution learning does not involve the prediction of pedestrian position, and we adopt the center map, scale map and offset map to generate the center point position and size of the pedestrian. The position confidence above 0.1 is defined as the positive center, and we combine the scale map to generate the pedestrian coordinates. We adopt NMS to filter redundant boxes, and the NMS is set as 0.5. Specifically, in the center prediction map, we define the above confidence score of 0.1 as the center of pedestrian, we take the index of the center position of the scale map and multiply it by 4 to obtain the center point of width and height. Add 0.5 to the value of the center point position in the offset map, and add the result in the center coordinate to obtain center point coordinates. Thus, we obtain the center position and scale of the pedestrian. We adopt a post-processing NMS threshold of 0.5 to filter prediction boxes.

4. Experiments

4.1. Experiment Setting

4.1.1. Datasets

CityPersons [11] Dataset was proposed in 2017, which is annotated on the CityScapes benchmark [22]. The background includes 27 cities. It has 5000 images altogether. We used 2975 images for training and 500 validation subset images for testing. The input scale of images is 1× when testing. The evaluation metric is MR^{-2} [23], which is the log–average Missing Rate over False Per Image (FPPI) ranging in $[10^{-2}, 100]$.

4.1.2. Training Details

CSPRS is realized in Pytorch [24]. ResNeSt101 [17] is the backbone for extract feature, which was pre-trained on ImageNet [25]. Adam [26] is adopted to optimize the network. In the training, we adopt moving average weights [27] to improve the results. A mini-batch contains five images in one GPU. The type of GPU is Tesla V100-SXM2 or Tesla P100-PCIE. The input resolution is 512 × 1024. The initial learning rate is set as 4×10^{-4}, which is unchanging during the training unless otherwise stated. The epoch of the training is 150. We choose the best performance within the 150 epochs.

4.2. Ablation Study

Inspired by super-resolution learning, we predict the total image and segmentation map of the total image. The result shows that this idea is helping to improve the result, the results are seen in the first line of Table 1, this achieves 43.63% on the heavy set. We propose that predicting the total image contains the background, which has many other objects, and thus we only predict the body region to make the detector focus on the body region. The result is shown in the second line of Table 1, and this is better than in the first line.

Table 1. Comparison of different strategies. R stands for resolution learning branch. S stands for segmentation learning branch.

Region	R	S	Visible	Reasonable	Bare	Partial	Heavy
total image	+	+	–	10.99%	7.83%	9.89%	43.63%
pedestrian area	+	+	+	9.71%	5.83%	8.89%	42.53%

We only add a branch on the basis of CSP [2], and the result shows that this branch helps the detector to obtain better performance with the different GPU types and sizes as discussed in the ACSP [28] paper. In ACSP, the ablation study indicates the type and size of GPU to influence the performance. In OCSP [29], they train a CSP on one GPU, and the result is not better than the CSP paper. Training CSPRS on one GPU with five batch sizes.

We add a segmentation branch to the CSP, we first only add a segmentation branch to CSP, and the results are seen in Table 2 second row. We both add the segmentation branch and the resolution branch in the third row of Table 2. The result shows that, when both adding segmentation and resolution branch, the detector has a better result on the heavy set. In total, inspired by region resolution learning, we add a prediction branch to learn the input itself, which has one channel. In this way, the detector learns other backgrounds and other objects; therefore, we refine the region of resolution learning task, and the detector only learns the pedestrian region. In order to find a new balance, we attempt to add different styles of pedestrian extra features.

Table 2. Ablation study of different adding extra feature of segmentation and resolution and so on.

+Resolution	Segmentation	Visible	Reasonable	Bare	Partial	Heavy
+	–	–	9.35%	6.11%	8.44%	43.78%
–	+	–	9.84%	5.68%	9.50%	43.39%
+	+	+	9.71%	5.83%	8.89%	42.53%

4.3. Comparison with the State of the Arts

We compare the CSPRS with state-of-art methods in the validation subset in CityPersons datasets including FRCNN [11], FRCNN+Seg [11], TLL [30], ALF [31], OR-CNN [32], CSP [2], BCNet [15] and ACSP [28]. The results are shown in Table 3. We evaluate CSPRS on reasonable set, bare set, partial set and heavy set. All method are trained on the CityPersons [11] datasets without any extra data (except ImageNet [25]) and tested on a validation subset from the CityPersons datasets. In Table 3, we observe that CSPRS obtains MR^{-2} with 42.53% on the heavy set of the CityPersons datasets.

On a reasonable set, CSPRS obtains MR^{-2} with 9.71%. CSPRS achieves MR^{-2} of 5.83% on the bare set. On a partial set, it gains 8.89% MR^{-2}. It can be seen that CSPRS obtains a fine result without any occlusion strategies. Our proposed method CSPRS is based on CSP. Compare with CSP [2], from the Table 3, CSPRS improves by 1.29% on reasonable set; 6.77% on the heavy set; 1.91% on the partial set and 2.27% on the bare set. The ACSP [28] is the adaption of CSP. CSPRS improves MR^{-2} by 3.89% on the heavy set than ACSP [28]. It obtains slightly worse results on reasonable, partial and bare set; hence, we adopt a different experiment setting and compare with ACSP.

In ACSP [28], they attempted different GPU numbers and image number per GPU and found that two GPUs worked best with two images per GPU. The resolution of our input image is 0.5 times the size of the original image, and in ACSP [28], the resolution of the input image is the size of the original image. Our batch sizes are 5 with one GPU; however, in ACSP [28], its batch sizes are 2 with two GPUs. More importantly, we found that CSPRS could achieve better results in very early epochs in the training process compared with the same epochs for ACSP.

Table 3. Comparisons with state-of-the-art on the reasonable, heavy, partial and bare sets of the CityPersons datasets.

Method	Backbone	Reasonable	Heavy	Partial	Bare
FRCNN [11]	VGG-16	15.4%	-	-	-
FRCNN+Seg [11]	VGG-16	14.8%	-	-	-
TLL [30]	ResNet-50	15.5%	53.6%	17.2%	10.0%
ALF [31]	ResNet-50	12.0%	51.9%	11.4%	8.4%
OR-CNN [32]	VGG-16	12.8%	55.7%	15.3%	6.7%
CSP [2]	ResNet-50	11.0%	49.3%	10.8%	8.1%
BCNet [15]	ResNet-50	9.8%	53.3%	9.2%	5.8%
ACSP [28]	ResNet-101	9.3%	46.3%	8.7%	5.6%
CSPRS (ours)	ResNeSt-101	9.71%	42.53%	8.89%	5.83%

5. Conclusions

The high representative task guides the pedestrian detection task. Based on this, we proposed two ways to help the detector learn more pedestrian features. The region resolution learning task and segmentation learning task handle occlusion. We fused the region resolution learning task and segmentation task to the pedestrian detection field. These two high representation tasks guide pedestrian detection. In this way, the detector becomes more robust. We found that both region resolution learning and fine-grained segmentation learning helped to improve the performance.

After adding region resolution learning to pedestrian detection, CSPRS learned the resolution pixel details and structure of pedestrians, and this resolution learning focused on the region of the pedestrian. After adding region segmentation learning to pedestrian detection, this segmentation branch learned the different features of each person. Experiments were conducted on the CityPersons dataset [11], and we achieved state-of-the-art performance on the heavy set. We suggest that these two high-representation tasks contain pedestrian features from different angles, and thus it will be interesting to find better ways to balance these two high-representation tasks in the future.

Author Contributions: Y.Z. Ideal and experiments, Write and modify the manuscript, English proofing. H.W. Suggestions and Modifications the manuscript, Supervision, English proofing. Y.L. and M.L. English editing and correction. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Jilin Province Science and Technology Department Science and Technology Development Planning Project of China (20210101415JC), and the Jilin Province Education Department Scientific Research Planning Project of China (JJKH20210753KJ).

Data Availability Statement: The data are not publicly available.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study.

References

1. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv* **2015**, arXiv:1506.01497.
2. 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/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5187–5196.
3. Mao, J.; Xiao, T.; Jiang, Y.; Cao, Z. What can help pedestrian detection? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3127–3136.
4. Chi, C.; Zhang, S.; Xing, J.; Lei, Z.; Li, S.Z.; Zou, X. Pedhunter: Occlusion robust pedestrian detector in crowded scenes. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10639–10646.
5. Shang, M.; Xiang, D.; Wang, Z.; Zhou, E. V2F-Net: Explicit Decomposition of Occluded Pedestrian Detection. *arXiv* **2021**, arXiv:2104.03106.
6. 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), Milano, Italy, 10–15 January 2021; pp. 8453–8460.
7. Zhou, C.; Yuan, J. Bi-box regression for pedestrian detection and occlusion estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 135–151.
8. Lu, R.; Ma, H.; Wang, Y. Semantic head enhanced pedestrian detection in a crowd. *Neurocomputing* **2020**, *400*, 343–351. [CrossRef]
9. 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, Korea, 27 October–2 November 2019; pp. 9557–9566.
10. Wang, Y.; Zhu, C.; Yin, X.C. A Hybrid Self-Attention Model for Pedestrians Detection. In *Proceedings of the International Conference on Neural Information Processing*; Springer: Bangkok, Thailand, 2020; pp. 62–74.
11. 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. 3213–3221.
12. Hasan, I.; Liao, S.; Li, J.; Ullah Akram, S.; Shao, L. Pedestrian detection: The elephant in the room. *arXiv* **2020**, arXiv:2003.08799.
13. Daliparthi, V.S.S.A. The Ikshana Hypothesis of Human Scene Understanding. *arXiv* **2021**, arXiv:2101.10837.
14. Wang, L.; Li, D.; Zhu, Y.; Tian, L.; Shan, Y. Dual Super-Resolution Learning for Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3774–3783.
15. Sha, M.; Boukerche, A. Semantic Fusion-based Pedestrian Detection for Supporting Autonomous Vehicles. In Proceedings of the 2020 IEEE Symposium on Computers and Communications (ISCC), Rennes, France, 7–10 July 2020; pp. 1–6.
16. Cai, J.; Lee, F.; Yang, S.; Lin, C.; Chen, H.; Kotani, K.; Chen, Q. Pedestrian as Points: An Improved Anchor-Free Method for Center-Based Pedestrian Detection. *IEEE Access* **2020**, *8*, 179666–179677. [CrossRef]
17. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Lin, H.; Zhang, Z.; Sun, Y.; He, T.; Mueller, J.; Manmatha, R.; et al. Resnest: Split-attention networks. *arXiv* **2020**, arXiv:2004.08955.
18. Meletis, P.; Wen, X.; Lu, C.; de Geus, D.; Dubbelman, G. Cityscapes-Panoptic-Parts and PASCAL-Panoptic-Parts datasets for Scene Understanding. *arXiv* **2020**, arXiv:2004.07944.
19. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
20. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
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. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
23. 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. [CrossRef] [PubMed]
24. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in pytorch. In Proceedings of the 31st Conference on Neural Information Processing System, Long Beach, CA, USA, 4–9 December 2017.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
26. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
27. Tarvainen, A.; Valpola, H. Weight-averaged, consistency targets improve semi-supervised deep learning results. *arXiv* **2017**, arXiv:1703.01780.
28. Wang, W. Adapted Center and Scale Prediction: More Stable and More Accurate. *arXiv* **2020**, arXiv:2002.09053.
29. Wang, H.; Zhang, Y.; Ke, H.; Wei, N.; Xu, Z. Semantic Structural and Occlusive Feature Fusion for Pedestrian Detection. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 293–307.

30. 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 (ECCV), Munich, Germany, 8–14 September 2018; pp. 536–551.
31. 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 (ECCV), Munich, Germany, 8–14 September 2018; pp. 618–634.
32. 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.