2021 IEEE International Joint Conference on Biometrics (IJCB) | 978-1-6654-3780-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCB52358.2021.9484371

SADet: Learning An Efficient and Accurate Pedestrian Detector

Chubin Zhuang[1] Zongzhao Li[1,2] Xiangyu Zhu[1,2] Zhen Lei[1,2,3*] Stan Z. Li[4]
[1]CBSR & NLPR, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences
[4]School of Engineering, Westlake University
{chubin.zhuang,xiangyu.zhu,zlei,szli}@nlpr.ia.ac.cn {lizongzhao2020}@ia.ac.cn

Abstract

Although the anchor-based detectors have taken a big step forward in pedestrian detection, the overall performance of algorithm still needs further improvement for practical applications, e.g.*, a good trade-off between the accuracy and efficiency. To this end, this paper proposes a series of systematic optimization strategies for the detection pipeline of one-stage detector, forming a single shot anchor-based detector (SADet) for efficient and accurate pedestrian detection, which includes three main improvements. Firstly, we optimize the sample generation process by assigning soft labels to the outlier samples to generate semi-positive samples with continuous tag value between* 0 *and* 1*. Secondly, a novel Center-*IoU *loss is applied as a new regression loss for bounding box regression, which not only retains the good characteristics of IoU loss, but also solves some defects of it. Thirdly, we also design Cosine-NMS for the post-processing of predicted bounding boxes, and further propose adaptive anchor matching to enable the model to adaptively match the anchor boxes to full or visible bounding boxes according to the degree of occlusion. Though structurally simple, it presents state-of-the-art result and real-time speed of* 20 *FPS for VGA-resolution images (*640×480*) tested on one GeForce GTX 1080Ti GPU on challenging pedestrian detection benchmarks, i.e., CityPersons, Caltech, and human detection benchmark CrowdHuman, leading to a new attractive pedestrian detector.*

1. Introduction

Pedestrian detection is a fundamental and essential step for many pedestrian related applications, *e.g.*, human gait recognition [1], person re-identification [32], and is of great requirement to both accuracy and efficiency. Recent years have witnessed the remarkable success achieved by convolutional neural network (CNN) [13], which also inspires pedestrian detection. R-CNN [9] firstly applies CNN to object detection based on proposals generated by Selective Search [28]. Following R-CNN, Region Proposal Network (RPN) integrated with pre-defined anchors are designed to generate candidate proposals in a unified framework, forming the two-stage detector Faster R-CNN [22], which is the originator of anchor-based methods. Nevertheless, these two-stage anchor-based detectors are still far from practical applications for the low run-time efficiency.

Alternatively, aiming at higher run-time efficiency, the one-stage method (*e.g.*, SSD [16]) discards the second stage of Faster R-CNN and directly detects objects by regular and dense sampling over locations, scales and aspect ratios. Though faster, one-stage detector has not presented competitive results on common pedestrian detection benchmarks (*e.g.*, CityPersons [35] and Caltech [8]).

To pursue a better balance between accuracy and efficiency for the anchor-based detector, this paper proposes a series of improvements to the detection pipeline of pedestrian detector, which can be summarized as follows: **1)** We optimize the sample generation process by assigning soft labels ranging from 0 to 1 to these outlier samples according to their jaccard overlap, and further add these semi-positive samples to the training objective as well. In this way, not only more valid samples are generated, but also the robustness of detector is strengthened. **2)** A novel Center-IoU loss is specially designed as a new regression loss for more precise pedestrian localization, which not only retains the good characteristics of IoU loss, but also solves some defects of it. **3)** We design Cosine-NMS for the postprocess of predicted bounding boxes to reduce false detections of the adjacent overlapping pedestrians, and further propose adaptive anchor matching algorithm to enable the model to adaptively match the anchor boxes to full or visible bounding boxes according to the degree of occlusion, making

[*] Corresponding author.

978-1-6654-3780-6/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on February 01,2023 at 14:42:52 UTC from IEEE Xplore. Restrictions apply.

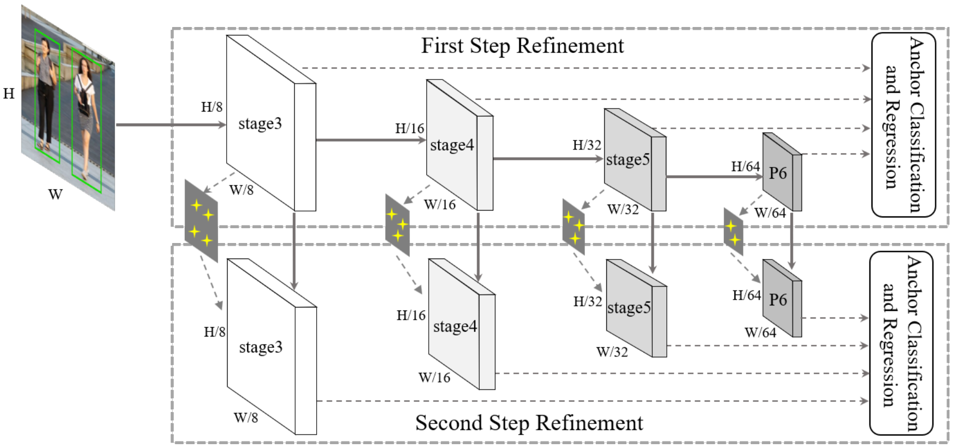

Figure 1. Architecture of SADet. For better visualization, we only display the layers used for detection. The celadon parallelograms denote the refined anchors associated with different feature layers after the first-step regression. The stars represent the centers of the refined anchor boxes, which are not regularly paved on the image.

the Non-Maximum Suppression (NMS) and anchor matching algorithms more suitable for occluded pedestrian detection. To better verify the effectiveness of the above methods, the relevant experiments are directly conducted on a very strong baseline ALFNet [17], which owns great performance on pedestrian detection. Notably, the aforementioned optimization strategies are systematic improvements to the detection pipeline of one-stage pedestrian detector, which can also be applied to any anchor-based pedestrian detector, including one-stage and two-stage methods.

In conclusion, this paper proposes a single shot anchor-based pedestrian detector (SADet) to achieve efficient and accurate pedestrian detection. Though structurally simple, the proposed SADet presents state-of-the-art result and real-time speed of 20 FPS for VGA-resolution images (640 × 480) on challenging pedestrian detection benchmarks, *i.e.*, CityPersons [35], Caltech [8], and human detection benchmark CrowdHuman [24], which demonstrates the supreme generalization capacity and effectiveness of our method.

2. Related work

Traditional detector. The traditional solution to this problem is training a discriminative pedestrian detector that exhaustively operates on the sub-images across all locations and scales. ACF [6], LDCF [21] extend the paradigm of Viola and Jones [29] to exploit various filters on Integral Channel Features (ICF) [7] with the sliding window strategy.

Anchor-based detector. Afterwards, coupled with the prevalence of deep learning techniques, the anchor-based methods originated from Faster R-CNN [22] rapidly dominate this field. As the pioneering work of anchor-based methods, Faster R-CNN [22] generates candidate proposals and further classifies and refines these proposals in a two-step regression framework. In contrast, one-stage detectors, popularized by SSD [16], remove the proposal generation step and achieve a great balance between accuracy and efficiency. ALFNet [17] stacks a series of predictors to directly evolve the default anchor boxes of SSD step by step into improving the detection results. In terms of pedestrian detection, the anchor-based detectors dominate.

Anchor-free detector. Considering the tedious design of anchors, the anchor-free detectors bypass this stage and directly make predictions on an image. DenseBox [11] first proposes a unified end-to-end fully convolutional framework that directly predicts bounding boxes. TLL [25] proposes to detect an object by predicting the top and bottom vertexes, which achieves significant improvement on Caltech [8]. CSP [18] simplifies pedestrian detection as a straightforward center and scale prediction task through convolutions and presents great performance on pedestrian detection benchmarks CityPersons [35] and Caltech [8].

3. Approach

This section introduces the details of the SADet that enable the detector to be accurate and efficient on pedestrian detection, including detection pipeline, anchor design, soft label design, Center-IoU loss, Cosine-NMS and adaptive anchor matching, as well as some other implementation details for training and inference.

3.1. Detection pipeline

In anchor-based detectors, multiple feature maps with different resolutions are extracted from a backbone network, which can be defined as follows:

$$\Phi_n = f_n(\Phi_{n-1}) = f_n(f_{n-1}(...f_1(I))), \quad (1)$$

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on February 01,2023 at 14:42:52 UTC from IEEE Xplore. Restrictions apply.

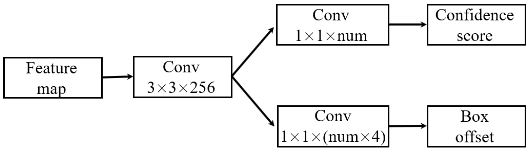

Figure 2. The detailed implementation of the detection module applied in our SADet, which is attached to each level of feature maps to translate default anchor boxes to corresponding detection results.

where I is the input image, $f_n(.)$ represents the n th layer and Φ_n is the generated feature map. On top of these multi-scale feature maps, detection can be formulated as:

$$p_n(\Phi_n, B_n) = \{cls_n(\Phi_n, B_n), reg_n(\Phi_n, B_n)\}, \quad (2)$$

where B_n is the anchor box pre-defined in the detection layer, $p_n(.)$ is the detection result of n th feature map, which consists of two elements, the classification scores $cls_n(\Phi_n, B_n)$ and the related parameters for anchor box regression. Following [17], we stack a series of predictors $p_n^t(.)$ on these anchor boxes to construct a two-step refinement framework for anchor boxes to improve the performance of one-stage detectors, which is depicted in Figure 1. Therefore, the above equation can be re-formulated as:

$$p_n(\Phi_n, B_n^0) = p_n^2(p_n^1(\Phi_n, B_n^0)), \quad (3)$$

This detection pipeline contains a two-step refinement of anchor boxes, the first step is used to coarsely adjust the locations and sizes of anchors to provide better initialization for the subsequent regressor. Then the second step takes the refined anchors as the input from the former to further improve the regression and classification process.

For the detection module, a set of convolutional layers are deployed to extract features from detection layers for pedestrian/non-pedestrian classification and bounding box regression, as depicted in Figure 2.

3.2. Anchor design

Anchor-based object detectors with reasonable design of anchors in different feature maps have proven to be effective to handle objects with different scales [37]. Our SADet applies ALFNet [17] as a baseline and uses ResNet-50 [10] as the backbone network of the detector, which is pictorially illustrated in Figure 1.

Table 1 presents the detailed anchor design of our SADet. The last layers of **stage3**, **stage4**, **stage5** in ResNet-50 and an additional convolutional layer **P6** attached at the end are selected as the detection layers with sizes downsampled by 8, 16, 32, 64 respectively, which are then associated with anchor boxes with width of (16, 24), (32, 48), (64, 96), (128, 160) pixels and a single aspect ratio of 0.41. Notably, the above settings of anchor boxes will be generally applied to all datasets without further adjustments in our experiments.

Table 1. The stride size, width and aspect ratio of pre-defined anchors of the four detection layers.

Detection Layer	Stride	Width	Aspect Ratio
stage3	8	16, 24	0.41
stage4	16	32, 48	0.41
stage5	32	64, 96	0.41
P6	64	128, 160	0.41

3.3. Soft label design

In common paradigm of sample generation process of anchor-based methods, two threshold values $\{T_{neg}, T_{pos}\}$ are usually pre-defined. During anchor matching phase, the anchors with IoU (Intersection-over-Union) ratio higher than T_{pos} will be assigned a positive label 1, and a negative label 0 will be sent to an anchor if its IoU ratio is lower than T_{neg}. Meanwhile, the anchors with IoU ratio between T_{neg} and T_{pos} will be ignored subsequently, which makes the model sensitive to the pre-defined thresholds and cannot fully utilize all valid samples.

Therefore, we propose a new design of soft label to utilize these ignored anchors. The assignment of soft label is defined as follows.

$$label = \begin{cases} 0, & \text{if} \quad iou < T_{neg} \\ 1, & \text{if} \quad iou > T_{pos} \\ \frac{iou - T_{neg}}{T_{pos} - T_{neg}}, & others \end{cases} \quad (4)$$

where an anchor with iou higher than T_{neg} and lower than T_{pos} will be assigned a soft label between 0 and 1 in accordance with their IoU ratio.

We define these samples with soft label as semi-positive samples. Strengthening the use of these samples can not only increase the number of samples available, but also enhance the robustness of the model for labels.

3.4. Center-IoU loss

IoU is the most popular evaluation metric used in the object detection benchmarks. However, as illustrated in [23], there exists a gap between optimizing the commonly used distance losses for regressing the parameters of a bounding box and maximizing this metric value. To alleviate the aforementioned problem, improved version such as $GIoU$ (Generalized Intersection over Union) loss [23] is proposed as the bounding box regression loss functions. Although the $GIoU$ loss has achieved great improvements, some problems still exist. As illustrated in Figure 3, the red rectangular box represents the ground-truth box, the green and blue dotted boxes are two different prediction results. On

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on February 01,2023 at 14:42:52 UTC from IEEE Xplore. Restrictions apply.

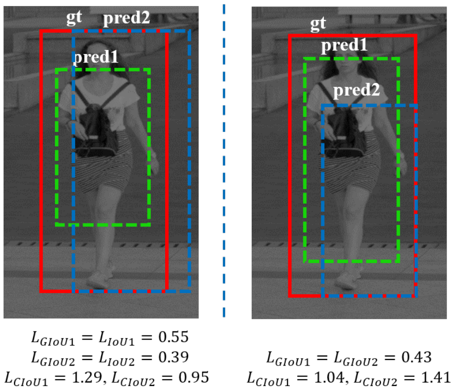
gt
pred2
pred1
gt
pred1
pred2
$L_{GIoU1} = L_{IoU1} = 0.55$
$L_{GIoU2} = L_{IoU2} = 0.39$
$L_{CIoU1} = 1.29, L_{CIoU2} = 0.95$
$L_{GIoU1} = L_{GIoU2} = 0.43$
$L_{CIoU1} = 1.04, L_{CIoU2} = 1.41$

Figure 3. Some hard examples for $GIoU$ loss. The box in red represents the ground-truth box, and the boxes in green and blue are two different prediction results.

one hand, the left image depicts that if the predicted box is completely enclosed by the ground-truth box, or these two boxes are in parallel state, the $GIoU$ loss will then degenerate into a common IoU loss. On the other hand, for the right image in Figure 3, we can find that no matter how we move the position of the predicted box, the $GIoU$ loss is always a fixed value.

To handle these problems, we propose a new bounding box regression loss, named Center-IoU loss. The definition of this loss function is detailed in Equation. 5 and 6.

$$L_{CIoU} = \text{smooth}_{\ln}(\frac{|C \setminus (B_{gt} \cap B_{pred})|}{|C|}) + \text{smooth}_{\text{L1}}(t_i, t_i^*), \tag{5}$$

$$\text{smooth}_{\ln}(x) = \begin{cases} -\ln(1-x), & \text{if} \quad x \le \sigma, \\ \frac{x-\sigma}{1-\sigma} - \ln(1-\sigma), & \text{if} \quad x > \sigma, \end{cases} \tag{6}$$

where B_{gt} and B_{pred} are the ground-truth and predicted bounding boxes, respectively. C is the smallest rectangular box enclosing both B_{gt} and B_{pred}. $|C \setminus (B_{gt} \cap B_{pred})|$ represents the volume (area) occupied by C excluding the intersection of B_{gt} and B_{pred}. $\text{smooth}_{\ln}(x)$ [31] is a smoothed $\ln(x)$ function which is continuously differentiable in (0, 1), and $\sigma \in$ [0, 1) is the smooth parameter to adjust the sensitiveness of the regression loss to the outliers. t_i and t_i^* are vectors representing the 2 parameterized center point coordinates of the predicted and ground-truth box, respectively. The distance regression term $\text{smooth}_{\text{L1}}(t_i, t_i^*)$ is the robust distance loss function defined in [30]. As shown in Figure 3, our Center-IoU loss can easily handle the hard examples for $GIoU$ loss, which proves the superiority of our method.

The proposed Center-IoU loss not only retains the good characteristics of $GIoU$ loss, but also solves some defects of it, enabling the regression process more suitable for pedestrian detection.

3.5. Cosine-NMS

Non-Maximum Suppression (NMS) is an integral part of the object detection pipeline, which recursively selects the detection box with the maximum score and remove the repeated predictions. However, applying NMS with a low threshold like 0.3 may increase the miss-rate, especially in crowd scenes. On the contrary, a high threshold like 0.6 may also increase false positives. Therefore, to better detect occluded pedestrians, we propose a novel Cosine-NMS for the postprocess of predictions based on the design of Soft-NMS [2]. The pruning step of this algorithm can be written as a re-scoring function as follows.

$$s_i = \begin{cases} s_i, & \text{iou}(M, b_i) < N_t, \\ s_i f(\text{iou}(M, b_i)), & \text{iou}(M, b_i) \ge N_t, \end{cases} \tag{7}$$

where N_t is the pre-defined threshold, $f(\text{iou}(M, b_i))$ is an overlap based weighting function to change the classification score s_i of a box b_i which has a high overlap with M. As for the design of weighting function, we propose a new weighting function based on cosine function to make the NMS algorithm more suitable for pedestrian detection, which is detailed as follows.

$$f(\text{iou}(M, b_i)) = \cos(\frac{\pi}{2}(\text{iou}(M, b_i) - N_t)/(1 - N_t)), \tag{8}$$

In soft-NMS, the Linear version $f(\text{iou}(M, b_i)) = (1 - \text{iou}(M, b_i))$ is not continuous in terms of overlap with a fixed gradient and a sudden penalty is applied when a NMS threshold of N_t is reached. Meanwhile, the Gaussian version $f(\text{iou}(M, b_i)) = e^{-\frac{\text{iou}(M, b_i)^2}{\sigma}}$ is continuous in terms of overlap, but the penalty term is introduced to any prediction box overlaps with ground-truth box and is too small for the highly overlapped boxes.

For comparison, the proposed Cosine penalty function for NMS is continuous in terms of overlap when the IoU ratio is higher than threshold N_t. For the detection boxes with IoU ratio lower than N_t, they are more likely to be true positives and thus no processing is done to them, otherwise a penalty term will be applied to decay their scores. Notably, the proposed Cosine-NMS has no additional parameters and can be directly applied to the post-processing of any pedestrian detectors.

3.6. Adaptive anchor matching

In common paradigm of anchor matching, the IoU ratios are usually calculated between pre-defined anchor boxes and full ground-truth bounding boxes based on the jaccard overlap. However, for the occluded pedestrian, the visible bounding box for some instance is only part of full bounding box, which contains a lot of redundant background information and makes it invalid to directly match the anchor

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on February 01,2023 at 14:42:52 UTC from IEEE Xplore. Restrictions apply.

to the full bounding box. To obtain a better result, we propose adaptive anchor matching to optimize this process as follows.

1) For ground-truth boxes with visible ratio R_{vis} lower than threshold T_{vis}, apply the visible bounding boxes to match the anchor boxes to reduce the interference of background information. The visible ratio R_{vis} is defined as the ratio of the visible bounding box and full bounding box, *i.e.* $area(B_{\text{vis}})/area(B_{\text{full}})$.

2) For ground-truth boxes with visible ratio R_{vis} higher than threshold T_{vis}, directly apply the full bounding boxes to match the anchor boxes to make full use of the context information to assist pedestrian detection.

3.7. Other implementations

Training. Anchors are assigned as positives S_+ if the $IoUs$ with any ground truth are above a threshold T_{pos}, and negatives S_- if the $IoUs$ lower than a threshold T_{neg}. Those anchors with IoU in $[T_{neg}, T_{pos})$ are also assigned as semi-positives S_* with label value ranging from 0 to 1 and sent to the training objective as well, which is detailed in Section. 3.3.

At each regression step t, the convolutional predictor is optimized by a multi-task loss function combining two objectives, which is illustrated as follows.

$$L(p,x) = \frac{1}{N_{cls}} \sum_{i=1} L_{cls}(p_i, p_i^*) + \frac{\lambda}{N_{reg}} \sum_{i=1} [p_i^* > 0] L_{reg}(x_i, x_i^*), \tag{9}$$

where i is the index of an anchor, p_i and p_i^* are the predicted probability and ground-truth label of anchor i. x_i is a vector representing the 4 parameterized coordinates of the predicted locations of pedestrian, while x_i^* is the corresponding ground-truth box parameters associated with a non-negative anchor. The regression loss L_{reg} is the proposed Center-IoU loss detailed in Section. 3.4. The classification loss $L_{cls}(p_i, p_i^*)$ is an optimized Focal Loss [14] for positive, negative and semi-positive samples, which is formulated as:

$$\begin{aligned} L_{cls}(p_i, p_i^*) = &-\alpha \sum_{i \in S_+} (1-p_i)^{\gamma} \log(p_i) - \beta \sum_{i \in S_*} (p_i^*)^{\gamma} \log(p_i) \\ &- (1-\alpha) \sum_{i \in S_-} p_i^{\gamma} \log(1-p_i), \end{aligned} \tag{10}$$

where α and γ are the focusing parameters, which are set to 0.25 and 2 as suggested in [14]. The first and third terms in loss function is the traditional Focal Loss, while the second term is the specially designed loss function for semi-positive samples with balancing parameter β fixed to 0.1, which assigns greater weights to the samples with higher label value.The cls term is normalized by the number of positive, negative and semi-positive anchors, and the reg term is normalized by the number of non-negative anchors.

Table 2. Ablative results on CityPersons validation set. MR^{-2} is used to compare the performance of detectors (lower score indicates better performance). The top result is highlighted in red.

Component	**SADet**				
Center-IoU		√	√	√	√
Soft Label			√	√	√
Cosine-NMS				√	√
Adaptive					√
Reasonable	**16.0**	**15.2**	**13.4**	**12.9**	**11.5**
Bare	11.3	10.0	8.7	8.2	6.7
Partial	14.8	14.2	12.9	12.1	10.7
Heavy	52.5	52.0	51.7	51.5	56.4

Data augmentation. To increase the robustness of training data, each training image is sequentially processed by color distortion, horizontal flipping and random crop. The resolutions of final training image are 336*448, 640*1280 and 832*832 for Caltech, CityPersons and CrowdHuman, respectively.

Inference. SADet simply involves feeding forward an image through the network. For each level, we get the regressed anchor boxes from the final predictor and hybrid confidence scores from all predictors. We first filter out most boxes by a confidence threshold of 0.05 and keep the top 1000 boxes before applying NMS, then all remaining boxes are merged with the proposed Cosine Non-Maximum Suppression (Cosine-NMS) with a threshold of 0.3, and the top 150 boxes will be selected as the output.

4. Experiments

4.1. Experimental setup

Our method is implemented in the Keras [3] library, with 2 GTX 1080Ti GPUs for model training. A mini-batch contains 15 images per GPU. For CityPersons [35], the backbone network is pre-trained on ImageNet [5]. The network is totally trained for 120 epoches, with the initial learning rate of $1e^{-3}$ and decreased by a factor of 10 after 60 and 80 epoches. For Caltech [8], we also include experiments with the model initialized from CityPersons as done in [17] and totally trained for 100 epoches with the learning rate of $1e^{-5}$. For CrowdHuman [24], the network is totally trained for 140 epoches with the base learning rate $1e^{-3}$ and divided by 10 after 80 and 100 epoches. The backbone network is ResNet-50 unless otherwise stated.

4.2. CityPersons dataset

The CityPersons [35] dataset is built upon the semantic segmentation dataset Cityscapes to provide a new dataset

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on February 01,2023 at 14:42:52 UTC from IEEE Xplore. Restrictions apply.

Table 3. Comparison with the state-of-the-art methods on the CityPersons validation set. All models are trained on the training set. Detection results tested on the original image size (1024 × 2048 on CityPersons) are reported. MR^{-2} is used to compare the performance of detectors (lower score indicates better performance).

Method	Framework	Backbone	**Reasonable**	Bare	Partial	Heavy	Test Time
Adapted Faster RCNN [35]	Two-stage	VGG-16	15.4	-	-	-	-
OR-CNN [36]	Two-stage	VGG-16	12.8	6.7	15.3	55.7	-
Adaptive-NMS [15]	Two-stage	VGG-16	11.9	6.2	12.6	55.2	-
RNMS+PBM [12]	Two-stage	VGG-16	11.1	-	-	53.3	-
TLL+MRF [25]	Two-stage	ResNet-50	14.4	9.2	15.9	52.0	-
Repulsion Loss [31]	Two-stage	ResNet-50	13.2	7.6	16.8	56.9	-
DRNet [19]	Two-stage	ResNet-50	10.1	-	-	46.2	0.15s/img
ALFNet-2s [17]	One-stage	ResNet-50	12.0	8.4	11.4	51.9	0.27s/img
CSP(w/o offset) [18]	Anchor-free	ResNet-50	11.4	8.1	10.8	49.9	0.33s/img
CSP(with offset) [18]	Anchor-free	ResNet-50	11.0	7.3	10.4	49.3	0.33s/img
APD [34]	Anchor-free	ResNet-50	10.6	7.1	9.5	49.8	-
SADet-1step	One-stage	ResNet-50	11.5	6.7	10.7	56.4	0.24s/img
SADet-2step	One-stage	ResNet-50	**9.7**	5.7	9.8	52.8	0.26s/img

of interest for pedestrian detection, which contains $5,000$ images ($2,975$ for training, 500 for validation, and $1,525$ for testing) with about $35,000$ manually annotated persons plus $13,000$ ignore region annotations.

Following the evaluation protocol in CityPersons, we train our detector on the training set, and evaluate it on the validation sets. The log miss rate averaged over the false positive per image (FPPI) range of [10^{-2}, 1] (MR^{-2}) is used to measure the detection performance (lower score indicates better performance).

4.2.1 Ablation study

To have a better understanding of how each proposed component affects the final performance, we construct four variants and evaluate them on CityPersons validation dataset, shown in Table 2. Our baseline model is the same as ALFNet [17].

Center-IoU Loss. Firstly, the smooth L_1 loss for bounding box regression applied in ALFNet is replaced with the novel Center-IoU loss, which helps promote the accuracy of localization and contributes to the reduce of MR^{-2} by 0.8% on *Reasonable* subset. Besides, the improvement still holds for pedestrians with different degrees of occlusion, including *Bare* (occlusion $\leq 10\%$), *Partial* ($10\% <$ occlusion $< 35\%$) and *Heavy* ($35\% \geq$ occlusion) subsets.

Soft Label Design. Secondly, we incorporate the soft label design to sample generation process to fully utilize all valid samples and smooth the predictions of the model. The comparison between the third and fourth columns in Table 2 indicates that our soft label design effectively improves the performance, especially for *Reasonable* subset with MR^{-2} reduced from 15.2% to 13.4%.

Cosine-NMS. Thirdly, we substitute the proposed Cosine-NMS for the commonly used greedy-NMS to make the Non-Maximum Suppression algorithm more suitable for pedestrian detection. The result in Table 2 validates the effectiveness of Cosine-NMS, with MR^{-2} decreased by 0.5%, 0.5%, 0.8% and 0.2% on *Reasonable*, *Bare*, *Partial* and *Heavy* subsets.

Adaptive Anchor Matching. The last contribution of SADet is the proposed adaptive anchor matching, which deals with the the problem of poor matching between anchor boxes and ground-truth boxes in crowded scenes. As reported in Table 2, the improvements on *Reasonable*, *Bare*, *Partial* subsets are 1.4%, 1.5% and 1.4% respectively.

4.2.2 Evaluation results

We compare the proposed SADet with the state-of-the-art detectors on CityPersons in Table 3. Notably, SADet-$n$step represents the model with n steps refinement to anchor boxes.

Without any additional supervision like semantic labels or auxiliary regression loss, our SADet achieves state-of-the-art results on the validation set of CityPersons by reducing 0.9% MR^{-2} with $\times 1$ scale, surpassing all published anchor-based and anchor-free methods, which demonstrates the superiority of the proposed method in pedestrian detection.

4.3. Caltech dataset

The Caltech [8] is one of the most popular and challenging datasets for pedestrian detection. We use the new high quality annotations provided by [35] to evaluate the proposed method. The training and testing sets contains $42,782$ and $4,024$ frames, respectively.

Similar to [31], we evaluate the SADet on the *Reasonable* subset of the Caltech dataset, and compare it to other state-of-the-art methods in Figure 4. As shown in Figure 4,

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on February 01,2023 at 14:42:52 UTC from IEEE Xplore. Restrictions apply.

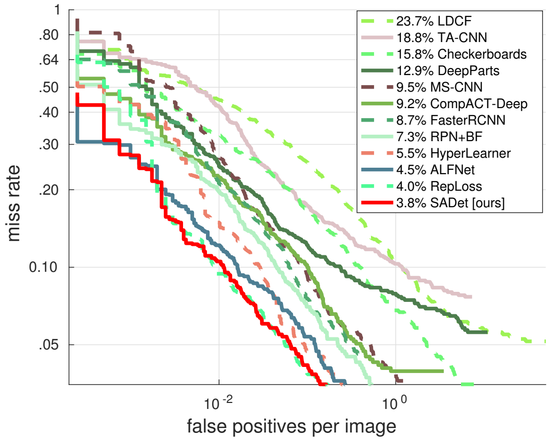

Figure 4. Comparisons with the state-of-the-art methods on the Caltech dataset. The scores in the legend are the MR^{-2} scores of the corresponding methods.

Table 4. Evaluation of full body detections on the CrowdHuman validation set.

Method	MR^{-2}	Recall	AP
RetinaNet [14]	63.33	93.80	80.83
ALFNet [17]	64.37	92.10	80.13
RFBNet-adaptive [15]	63.03	94.77	79.67
SADet-1step	62.96	94.22	80.01
SADet-2step	**60.13**	**95.02**	**82.14**

the proposed SADet performs competitively with the state-of-the-art result [4, 17, 20, 26, 27, 31, 33] on the Caltech dataset, surpassing all one-stage and two-stage detectors.

4.4. CrowdHuman dataset

Given that the size of CityPersons [35] and Caltech [8] datasets are not particularly large, we further carry out experiments on the newly released CrowdHuman [24] dataset to validate the generalization capacity of the proposed method.

The results of the proposed SADet and other state-of-the-art methods on CrowdHuman dataset are reported in Table 4. Based on the two-step regression structure, our SADet outperforms all the other one-stage detectors with 60.13% MR^{-2}, 95.02% Recall and 82.14% AP on CrowdHuman, which not only demonstrates the superiority of our method, but also verifies its generalization capacity to other scenarios, *i.e.*, human detection.

5. Conclusion

In this paper, we propose a series of systematic optimization strategies for the detection pipeline of one-stage detector, forming a single shot anchor-based detector (SADet) for efficient and accurate pedestrian detection. Specifically, we first introduce a new design of soft label to the sample generation process to make full use of all valid samples and improve the robustness of classification. Then we propose a new bounding box regression loss, named Center-IoU loss, which not only alleviates the defects of $GIoU$ loss, but also enforces the predicted bounding boxes to be close to the associated objects and locate compactly. Meanwhile, for occluded pedestrian detection, we design Cosine-NMS for the postprocess of predictions to reduce false detections of the adjacent overlapping pedestrians, by assigning a higher penalty to the highly overlapped detections. Besides, the adaptive anchor matching is also proposed to enable the model to adaptively match the anchor boxes to full or visible bounding boxes according to the degree of occlusion. Our method is trained in an end-to-end fashion and achieves state-of-the-art accuracy on challenging pedestrian detection benchmarks, *i.e.*, CityPersons [35], Caltech [8], and human detection benchmark CrowdHuman [24] with real-time speed of 20 FPS for VGA-resolution images.

Acknowledgement

This work was supported by the National Key Research and Development Program of China (No. 2020YFC2003901).

References

[1] H. Arshad, M. A. Khan, M. Sharif, M. Yasmin, and M. Y. Javed. Multi-level features fusion and selection for human gait recognition: an optimized framework of bayesian model and binomial distribution. *International Journal of Machine Learning and Cybernetics*, 10(12):3601–3618, 2019.

[2] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. Improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017.

[3] F. Chollet et al. Keras: Deep learning library for theano and tensorflow. *URL: https://keras. io/k*, 7(8):T1, 2015.

[4] A. Daniel Costea and S. Nedevschi. Semantic channels for fast pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2360–2368, 2016.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1532–1545, 2014.

[7] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. 2009.

[8] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on February 01,2023 at 14:42:52 UTC from IEEE Xplore. Restrictions apply.

segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] L. Huang, Y. Yang, Y. Deng, and Y. Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015.

[12] X. Huang, Z. Ge, Z. Jie, and O. Yoshie. Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10750–10759, 2020.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[15] S. Liu, D. Huang, and Y. Wang. Adaptive nms: Refining pedestrian detection in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6459–6468, 2019.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[17] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *European Conference on Computer Vision*, pages 618–634, 2018.

[18] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2019.

[19] Y. Luo, C. Zhang, M. Zhao, H. Zhou, and J. Sun. Which to match? selecting consistent gt-proposal assignment for pedestrian detection. *arXiv preprint arXiv:2103.10091*, 2021.

[20] J. Mao, T. Xiao, Y. Jiang, and Z. Cao. What can help pedestrian detection? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3127–3136, 2017.

[21] W. Nam, P. Dollár, and J. H. Han. Local decorrelation for improved pedestrian detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2014.

[22] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[23] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[24] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

[25] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *European Conference on Computer Vision*, pages 536–551, 2018.

[26] Y. Tian, P. Luo, X. Wang, and X. Tang. Deep learning strong parts for pedestrian detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1904–1912, 2015.

[27] Y. Tian, P. Luo, X. Wang, and X. Tang. Pedestrian detection aided by deep learning semantic tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5079–5087, 2015.

[28] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[29] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[30] H. Wang, F. Nie, and H. Huang. Robust distance metric learning via simultaneous l1-norm minimization and maximization. In *International conference on machine learning*, pages 1836–1844, 2014.

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

[32] J. Wu, Y. Yang, H. Liu, S. Liao, Z. Lei, and S. Z. Li. Unsupervised graph association for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8321–8330, 2019.

[33] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Convolutional channel features. In *Proceedings of the IEEE international conference on computer vision*, pages 82–90, 2015.

[34] J. Zhang, L. Lin, Y. Li, Y.-c. Chen, J. Zhu, Y. Hu, and S. C. Hoi. Attribute-aware pedestrian detection in a crowd. *arXiv preprint arXiv:1910.09188*, 2019.

[35] S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017.

[36] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *European Conference on Computer Vision*, pages 637–653, 2018.

[37] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 192–201, 2017.

Authorized licensed use limited to: HUNAN UNIVERSITY. Downloaded on February 01,2023 at 14:42:52 UTC from IEEE Xplore. Restrictions apply.