

# Adapted Center and Scale Prediction: More Stable and More Accurate

Wenhai Wang <sup>\*1</sup>

<sup>1</sup>*School of Mathematical Sciences(SMS), Beihang University, Beijing, China*

## Abstract

Pedestrian detection benefits from deep learning technology and gains rapid development in recent years. Most of detectors follow general object detection frame, *i.e.* default boxes and two-stage process. Recently, anchor-free and one-stage detectors have been introduced into this area. However, their accuracies are unsatisfactory. Therefore, in order to enjoy the simplicity of anchor-free detectors and the accuracy of two-stage ones simultaneously, we propose some adaptations based on a detector, Center and Scale Prediction(CSP). The main contributions of our paper are: (1) We improve the robustness of CSP and make it easier to train. (2) We propose a novel method to predict width, namely compressing width. (3) We achieve the second best performance on CityPersons benchmark, *i.e.* 9.3% log-average miss rate(MR) on reasonable set, 8.7% MR on partial set and 5.6% MR on bare set, which shows an anchor-free and one-stage detector can still have high accuracy. (4) We explore some capabilities of Switchable Normalization which are not mentioned in its original paper.

**Keywords:** Pedestrian Detection, Anchor-free, Switchable Normalization, Convolutional Neural Networks

## 1 Introduction

With the prevalence of artificial intelligence technique, autonomous vehicles have gained more and more attention. Although automatic driving needs integration of a lot of technologies, pedestrian detection is one of the most important. That's because missing pedestrian detection could threaten pedestrians' lives. As a result, the performance of pedestrian detection algorithms is of great importance.

With the development of generic object detection [8, 9, 22, 30–32], the detection performance on benchmark datasets [2, 6, 7, 35, 49] is significant improved. Also, some detectors, such as [23, 24, 28, 47], are specially designed for pedestrian detection.

However, though detection performance is improved on benchmark datasets all the time, there is still a huge gap between current pedestrian detector and a careful people [48]. Therefore, further performance improvement is necessary. Take pedestrian detection dataset, CityPersons [49], for instance. For a fair comparison, the following log-average miss rates(denoted as MR)(lower is better) are reported on the reasonable validation set with the same input scale (1x). From all of the state-of-the-arts literature available(including preprint ones), we summarize as follows: Repulsion Loss [43](13.2%), OR-CNN [52](12.8%), HBAN [25](12.5%), ALF [23](12.0%), Adaptive NMS [21](11.9%), CSP [24](11.0%), MGAN [28](10.5%), PSC-Net [45](10.4%), APD [47](8.8%). In the aforementioned state-of-the-arts methods, most of them have special occlusion/crowd handling process(7/9):

---

<sup>\*</sup>Corresponding author: wangwenhai@buaa.edu.cn

Repulsion Loss [43], OR-CNN [52], HBAN [25], Adaptive NMS [21], MGAN [28], PSC-Net [45], APD [47]. In addition, APD [47] uses more powerful backbone, *i.e.* DLA-34 [46], to improve MR from 10.6%(ResNet-50 [10]) to 8.8%. APD [47] also takes advantage of post process like Adaptive NMS [21].

For CSP [24], there is no special occlusion/crowd handling process or more powerful backbone. And it achieves competitive MR with other methods. However, there are also some unsolved problems existing in CSP [24]. First, it is sensitive to the batch size. More specifically, in the case of a small batch size, such as (1, 1)(The bracket (, ) denotes (#GPUs, #samples per GPU)), or a big batch size, such as (4, 4), MR will not converge, *i.e.* MR will approach 1 after several iterations. Second, when training CSP [24], different input scales bring significantly different performance. Finally, when compared to algorithms with occlusion/crowd handling process, there is still much room for improvement.

To address the above limitations, we propose **Adapted Center and Scale Prediction (ACSP)**, which has slight difference with original CSP [24] but brings significant improvement on performance. Detection examples using ACSP are shown in Fig. 1. In summary, the main contributions of this paper are as follows: (1) We make original CSP [24] more robust, *i.e.* less sensitive to batch size and input scale. (2) We propose compressing width, a novel method to determine the width of a bounding box. (3) We improve the performance of CSP [24]. (4) We explore the power of Switchable Normalization when the batch size is big.

Experiments are conducted on the CityPersons [49] database. We achieve the second best performance, *i.e.* 9.3% MR on reasonable set, 8.7% MR on partial set, 5.6% MR on bare set.

## 2 Related Works

### 2.1 Generic Object Detection

Early object detection approaches, such as [5, 19, 42], mainly utilize region proposal classification and sliding window paradigm. Since August 2018, more and

more works focus on anchor-free approaches. As a result, modern object detectors can be divided into two classes: anchor-based and anchor-free.

#### 2.1.1 Anchor-based

Anchor-based methods includes two-stage detectors and one-stage detectors. The most famous series of two-stage detectors are RCNN [9] and its descendants, *i.e.* Fast-RCNN [8], Faster-RCNN [32]. They build two-stage framework, which contains object proposals and classification. For one-stage detectors, YOLOv2 [31] and SSD [22] successfully accomplish detection and classification tasks on feature maps at the same time.

#### 2.1.2 Anchor-free

The earliest exploitation of anchor-free mode comes from DenseBox [12] and YOLOv1 [30]. The main difference between them is that DenseBox is designed for face detection while YOLOv1 is a generic object detection. The introduction of CornerNet [16] brings anchor-free detection into key point era. Its followers include ExtremeNet [56], CenterNet [55], etc. In addition, FoveaBox [15] and FSAF [57] use dense object detection strategy.

### 2.2 Pedestrian Detection

Before the dominance of deep learning techniques, traditional pedestrian detectors, such as [5, 27, 50], focus on integral channel features with sliding window strategy. Recently, with the introduction of Faster RCNN [32], some two-stage pedestrian detection approaches [21, 28, 43, 49, 51–54] achieve state-of-the-arts on standard benchmarks. Also, some pedestrian detectors [17, 21, 23], which base on single-stage backbone, gain a balance between speed and accuracy.

Zhou *et al.* [53] propose a discriminative feature transformation which enforces feature separability of pedestrian and non-pedestrian examples to handle occlusions for pedestrian detection. In [52], a new occlusion-aware R-CNN is proposed to improve the detection accuracy in the crowd. Wang *et al.* [43] develop a novel loss, repulsion loss, to address crowd



Figure 1: We use CityPersons test set to illustrate our ACSP detection ability. It is worthy to mention that, without any occlusion handling method, small and occlusion pedestrian can still be detected. The detections are shown in green rectangle boxes.

occlusion problem. The work of [21] focuses on Non-Maximum Suppression and proposes a dynamic suppression threshold to refine the bounding boxes given by detectors. HBAN [25] is introduced to enhance pedestrian detection by fully utilizing the human head prior. ALFNet is proposed in [23] to use asymptotic localization fitting strategy to evolve the default anchor boxes step by step into precise detection results. MGAN [28] emphasizes on visible pedestrian regions while suppressing the occluded ones by modulating full body features. PSC-Net [45] is designed for occluded pedestrian detection. CSP [24] utilizes an anchor-free method, *i.e.* directly predicting pedestrian center and scale through convolutions. Based on CSP [24], Zhang *et al.* propose APD [47] to differentiate individuals in crowds. All of the aforementioned methods achieve state-of-the-arts on CityPersons benchmark [49].

### 2.3 Normalization

Batch Normalization(BN) [13] is proposed to accelerate training process and improve the performance of CNNs. [34] points out that batch normalization makes the loss surface smoother while the original paper [13] believes the improvement comes from "internal covariate shift". Although, even today, it is still unknown that why batch normalization works so well, the utilization of batch normalization improves the performance of object detection, image classification, etc.

After batch normalization, weight normalization(WN) [33] is introduced to normalize the weights of layers. Layer normalization(LN) [1] normalizes the inputs across the features instead of the batch dimension. In this way, the performance will not be influenced by batch size and layer normalization is used in RNN at first. Originally designed for style transfer, instance normalization(IN) [41] normalizes across each channel in each training example. Group nor-

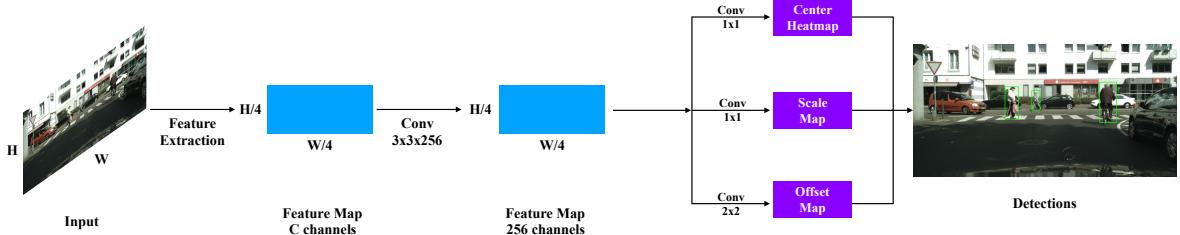


Figure 2: It is the architecture of original CSP [24]. The frame includes two parts: feature extraction and detection head.

malization(GN) [44] divides the channels into groups and computes the mean and variance for normalization within each group. As a result, it addresses the problem that, when the batch size becomes smaller, the performance of batch normalization goes down. It is a combination of layer normalization and instance normalization to some degree.

Recently, Luo *et al.* propose switchable normalization(SN) [26], which uses a weighted average of different mean and variance statistics from batch normalization, instance normalization, and layer normalization.

### 3 Proposed Adaptation

#### 3.1 CSP Revisit

CSP [24] was proposed by Wei Liu and Shengcai Liao in 2019. They first introduced anchor-free method into pedestrian detection area. More specifically, CSP [24] includes two parts: feature extraction and detection head. In feature extraction part, a backbone, such as ResNet-50 [10], MobileNet [11], is used to extract different levels of features. Shallower feature maps can provide more precise localization information while deeper feature maps can provide high-level semantic information. In detection head part, convolutions are used to predict center, scale, and offset respectively. In Fig. 2, we summarize the architecture of CSP [24].

A more detailed architecture of CSP [24] will be revisited in this paragraph. However, it will be slightly different with original paper [24]. We take ResNet-50

[10] and a picture with original shape, *i.e.*  $1024 \times 2048$  for instance. The difference between keeping original shape and resizing picture to  $640 \times 1280$  as showed in [24] will be discussed in ablation study. First, CSP [24] enlarges a picture with 3 channels into 64 channels through a  $7 \times 7$  Conv layer. Certainly, BN layer, *ReLU* layer and *Maxpool* layer follow the Conv layer. In this way, a  $(3, 1024, 2048)$  (The bracket  $(, , )$  denotes (#channels, height, width)) picture will be turned into a  $(64, 256, 512)$  one. Second, CSP [24] take 4 layers from ResNet-50 [10] with dilated convolutions. The 4 layers downsample the input image by 4, 8, 16, 16 respectively. At that time, we get 4 feature maps:  $(256, 256, 512)$ ,  $(512, 128, 256)$ ,  $(1024, 64, 128)$ ,  $(2048, 64, 128)$ . CSP [24] chooses to use a deconvolution layer to fuse the last 3 multi-scale feature maps into a single one. As a result, a  $(768, 256, 512)$  final feature map is made. Third, a  $3 \times 3$  Conv layer is used on the final feature map to reduce its channel dimensions to 256. Finally, three convolutions:  $1 \times 1$ ,  $1 \times 1$  and  $2 \times 2$  are appended for the prediction of center, scale and offset respectively.

#### 3.2 SN Layer

According to the aforementioned revisit, we conclude that there are totally 50 BN layers in CSP [24]. Although BN layer brings performance improvement to CSP [24] as it brings to other tasks, CSP [24] also suffers from the drawback of BN layer. On one hand, BN layer is unsuitable when the batch size is small. That is because small batch size will make the training process noisy, *i.e.* the amplitude of training loss

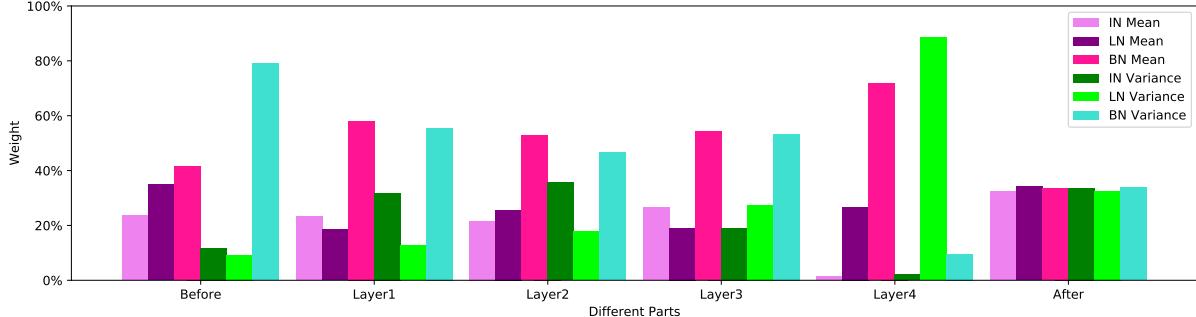


Figure 3: The proportion of the weight of each normalization method in different parts is shown in the histogram. The weights of mean and variance are displayed separately.

is relatively huge. However, ablation study will show a bigger batch size, even (1, 8), brings more harm to CSP [24]. More specifically, MR of validation set will decrease to near 16% and then increase to 1. It is likely that CSP [24] falls into local optimum and loses generalization ability.

To address this limitation, we replace all *BN* layers with *SN* layers. The effectiveness of this change will be shown in the ablation part, we try to explain the reason of it now.

To illustrate more specifically, we take (1, 8) for instance and the backbone is ResNet-50. The architecture of network can be divided into 6 parts: The first 5 come from backbone and the last one(denoted as After) is detection head. The first part(denoted as Before) is the operations before 4 layers in ResNet-50. The next four is the four layers. There is only 1 *BN* layer in the first part while there are 9, 12, 18, and 9 *BN* layers in the next four parts respectively. Finally, the detection head has only 1 *BN* layer. As suggested in 2.3, Switchable normalization is the combination of batch normalization, instance normalization, and layer normalization with different weights. Therefore, exploring the proportion of different weights in each part of network will show what makes a difference on earth. For each part, we calculate the weights of each normalization method in *SN* layers. Then the average of these weights are shown in Fig. 3. Although different normalization methods have different weights in each part, we figure out two

main differences with original *BN* layers. For one hand, in Layer4, the weight of *BN* variance is very small while the weight of *LN* variance is very big. For the other hand, in 'After' Part, *IN*, *LN* and *BN* share similar weights. The conclusions are as follows: (i) The low *BN* variance in Layer4 decreases the influence of noise when estimating variance. In this way, high-level semantic information can be utilized fully during inference process. (ii) The similar weights in 'After' part enable these three normalization methods to play same important roles. (iii) Different normalization methods in all parts complement one another.

### 3.3 Backbone

The feature extraction ability of backbone is of great importance in object detection. Some networks, such as, ResNet-50 [26], ResNet-101 [26], VGG [36] and MobileNet [11], which are original designed for image classification, are widely used in pedestrian detectors. In addition, some other networks, such as DetNet [18], are specially designed for object detection.

In the original paper [24], ResNet-50 [26] and MobileNet [11] are used as backbone. However, because of the nature of CSP [24], *i.e.* it fuses different level of feature maps, it is suitable to use deeper backbone network. In this way, the location information will still be stored in shallow feature maps and higher-level semantic information will be extracted at the same time.

Inspiring by the aforementioned idea, we select two

new backbones, expecting to obtain better performance. First, we use ResNet-101 [26] as our ACSP backbone. Compared to ResNet-50 [26], the only difference of ResNet-101 [26] is its third layer: there are 23 Bottleneck blocks rather than 6 Bottleneck blocks. As a result, in our ACSP, the last two feature maps presents higher level semantic information than CSP [24]. Meanwhile, localization information will not be changed. In theory, the fusion in our ACSP is more efficient than original CSP [24]. We will conduct ablation study to prove it. Second, in [18], authors point out that using DetNet [18] as backbone, they achieve state-of-the-art on the MSCOCO benchmark [20]. Therefore, it is likely that DetNet [18] will improve the performance of original CSP [24]. However, after fine tuning learning rate and so on, we find it is unpromising. We conclude the reason is that: one of the design concept of DetNet [18] is to address poor location problem, however, in CSP [24], this problem is solved by the fusion of different level layers and efficient center prediction.

### 3.4 Input Size

In the original paper [24], Liu *et al.* do not justify the resizing process, *i.e.* why in training part, the authors resize the original picture shape( $1024 \times 2048$ ) into  $640 \times 1280$ . After comparison, we find that: For one hand, resizing shape is beneficial to time-saving and memory-saving. More importantly, keeping original shape will worsen the performance of CSP [24] and bring non-convergent results. However, most of its counterparts take advantage of original resolution and achieve state-of-the-arts. Inspiring by this, we believe some adjustment will make a difference.

Based on the improvement in 3.2, we compare the performance between keeping and resizing. In ablation part, we will show that the performance also becomes worse, however. We conclude the reasons: First, some noise may exist in the original pictures. With the ResNet-50 backbone, the semantic features are not extracted adequately. Therefore, parameters of the network may be influenced by noise. Second, the quantity of parameter is not sufficient to fit the useful part of so high resolution pictures. Finally, resizing process will omit some detail features, and

focusing on them excessively will influence the ability of generalization.

To address the aforementioned problems, we replace ResNet-50 with ResNet-101 as suggested in 3.3. In this way, the performance is improved and we achieve the lowest MR of our ACSP. The reasons are as follows: For one hand, the increase of parameters enhances fitting ability of our ACSP. For the other hand, the increase of layers enables our ACSP to extract more high-level semantic feature and decrease the focus on details. Therefore, ACSP is immune to noise and has more generalization ability.

### 3.5 Compressing Width

From the original paper [24], we can see that the width of a box is obtained by multiplying the height by 0.41. It concurs with pedestrian aspect ratio in CityPersons Dataset [49]. However, it is not suitable in the reference process. That is because, in crowded scene, relatively wide boxes will increase the chance of overlapping and the NMS process will eliminate some of boxes. In this way, we will lose some detections.

As a result, we try to design a novel method to determine the width. On one hand, as we mentioned before, a wide box is not appropriate. On the other hand, a too narrow box is also not suitable. That is because, in this way, IoU between detections and ground truths will be small and detections will not be regarded as correct. Inspiring by the aforementioned analysis, we give our formula for calculating width:

$$w = r \cdot h,$$

where  $r$  is the aspect ratio( $r < 0.41$ ) and  $h$  is the predicted height of a bounding box.

It should be mentioned that the exact form of our compressing width is not crucial and we choose the most basic one. What matters most is the design concept.

### 3.6 Vanilla L1 Loss

As pointed in [24], total loss consists of classification loss, scale loss and offset loss. The weights are

0.01, 1 and 0.1, respectively. And for scale regression loss, [24] utilizes Smooth L1 to accelerate convergence. However, [38, 39, 55] show that vanilla L1 is better than Smooth L1. Therefore, we try to replace Smooth L1 with vanilla L1. We experimentally set the weights as 0.01, 0.05 and 0.1, respectively. The effectiveness of this improvement will be shown in ablation study.

## 4 Experiments

### 4.1 Experiment settings

#### 4.1.1 Dataset

To prove the efficacy of our adaptation, we conduct our experiments on CityPersons Dataset [49]. CityPersons is introduced recently and with high resolution. And the dataset is based on CityScapes benchmark [3]. It includes 5,000 images with various occlusion levels. We train our model on official training set with 2,975 images and test on the validation set with 500 images. In our test, the input scale is 1x.

#### 4.1.2 Training details

The ACSP is realized in Pytorch [29]. Adam [14] optimizer is utilized to optimize our network. Same as CSP [24] and APD [47], moving average weights [40] is adopted. Experiments show it helps achieve better performance. The backbone is fixed ResNet-101 [26] unless otherwise stated, *i.e.* replacing all *BN* layers with *SN* layers. It is pretrained on ImageNet [4]. We optimize the network on 2 GPUs (Tesla V100) with 2 images per GPU. The learning rate is  $2 \times 10^{-4}$  and training process is stopped after 150 epochs with 744 iterations per epoch. In the training process, we keep the original shape of pictures, *i.e.*  $1024 \times 2048$ .

### 4.2 Ablation Study

In this section, we conduct an ablative analysis on the CityPersons Dataset [49]. We use the most common and standard pedestrian detection criterion, log-average miss rates(denoted as MR), as evaluation

metric. In addition, the following MRs are all reported on reasonable set.

#### What is the influence of *SN* layer on stable training?

The stability of training process is of great importance. It comes from two aspects: whether the network is sensitive to the batch size and whether the performance will become poor after many iterations. To answer these two questions, we compare our ACSP with original CSP [24]. It should be mentioned that learning rate is appropriate in the following experiments, *i.e.* the training loss decreases and converges.

For the first one, comparisons are shown in Table 1. To conduct a fair comparison, the only difference is we replace all *BN* layers with *SN* layers, *i.e.* the backbone is still Resnet-50, the training input scale is still  $640 \times 1280$  and so on. In the table, the bracket (, ) denotes (#GPUs, #samples per GPU). For instance, (4, 4) means 4 GPUs with 4 images per GPU. 'Con' means the training is convergent, *i.e.* MR is still low no matter how many iterations are used. 'Exp' means the training is not convergent, *i.e.* MR increases to 1 after several iterations. The improvement line shows the percentage of decrease in MR from CSP [24] to ACSP. It is shown that when we choose GPU number and image number per GPU carefully, such as (4, 2), (2, 2), although ACSP outperforms CSP [24] to some degree, the improvement is not significant. However, when the batch size is bigger or smaller, such as (4, 4), (1, 1), ACSP brings conspicuous improvement. It is noteworthy that, though batch size is 8, there is a huge difference in MR between (4, 2) and (1, 8) for CSP [24]. That difference does not come from *BN* layer because *BN* layer will only be invalid when it is (8, 1) rather than (1, 8). Therefore, it is impossible to reproduce the reported result in [24] for someone who only has single GPU resource.

For the second one, we can come to a conclusion from Figure 4 and Table 1. For CSP [24], only (4, 2) and (2, 2) bring convergence result. However, for ACSP, all of the results are convergent.

#### How important is the backbone?

In this part, we compare three different backbones, *i.e.* ResNet-50 [26], ResNet-101 [26], DetNet [18]. The experiments are conducted based on *BN* layer

Table 1: Comparisons of different batch sizes and different methods. The bracket (, ) denotes (#GPUs, #samples per GPU). 'Con' means the training is convergent while 'Exp' means the training is not convergent. Bold number indicates the best result.

MR \ batch method	(4, 2)	(4, 4)	(2, 2)	(1, 1)	(1, 8)	(8, 1)
CSP	11.56% Con	27.75% Exp	11.34% Con	16.35% Exp	16.10% Exp	14.51% Exp
ACSP	11.16% Con	11.89% Con	<b>10.80%</b> Con	13.42% Con	11.66% Con	12.88% Con
Improvement	+3.46%	+57.15%	+4.76%	+17.92%	+27.58%	+11.23% Con

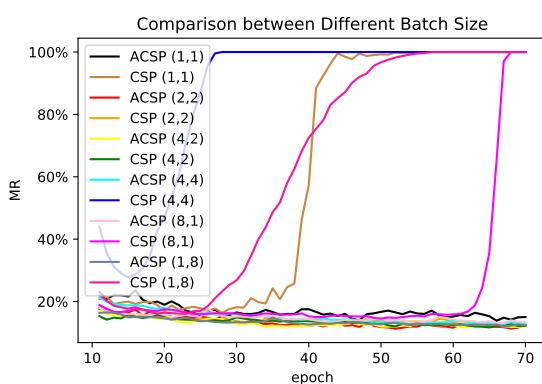


Figure 4: Comparisons of different batch size. It is shown that: For CSP [24], 4 experiment settings are not convergent; for ACSP, all experiments are convergent.

and *SN* layer respectively. The experiments setting is (4, 2). And the input size is  $640 \times 1280$ . The results are reported in Table 2.

We can conclude that: (i) As suggested in the theory part, ResNet-101 [26] outperforms ResNet-50 [26] no matter which normalization method is chosen. (ii) DetNet [18] underperforms ResNet-50 [26] and ResNet-101 [26] slightly. (iii) As discussed before, replacing *BN* layers with *SN* layers bring performance improvement on ResNet-50 [26] and ResNet-101 [26]. However, on DetNet [18], the MR increases. That is partly because we cannot find pretrained parameters of *SN* layers in DetNet [18].

### How important is the input resolution?

To prove the discussion in 3.4, we conduct some experiments with different resolutions under different circumstances. From Table 3, we can find that: For original CSP [24], the MR is not convergent when we do not resize pictures to  $640 \times 1280$ . When we use *SN*, as expected, the MR is convergent and performance is improved. But keeping original resolution is still not a better choice.

In Table 4, the experiments are conducted using *SN* as normalization method and ResNet-101 as backbone. As analysed in 3.4, the performance gets better no matter which batch size is chosen.

Table 2: Comparisons of different backbones and different normalization methods. Bold number indicates the best result. The experiments setting is (4, 2).

MR \ Method Backbone	BN	SN
ResNet-50	11.56%	11.16%
ResNet-101	11.29%	<b>10.91%</b>
DetNet	12.66%	12.91%

### What is the contribution of *SN* layer to the MR?

As stated in the before part, *SN* layer brings significant improvement when batch size is not carefully selected. In addition, from Table 2, we conclude *SN*

Table 3: Comparison between different resolutions under different normalization methods. Resolution part means the input picture scale. The experiments setting is (2, 2) and the backbone is ResNet-50.

MR \ Resolution Method	1024 × 2048	640 × 1280
BN	30.08%	11.34%
SN	11.41%	<b>10.80%</b>

Table 4: Comparison between different resolutions under different batch sizes. Resolution part means the input picture scale. The normalization method is SN and the backbone is ResNet-101.

MR \ Resolution Batch	1024 × 2048	640 × 1280
(2, 2)	<b>10.30%</b>	10.81%
(4, 2)	10.69%	10.91%

Table 5: Comparisons between different aspect ratio under different sets.

MR \ Set Ratio	Reasonable	Heavy	Partial	Bare
r = 0.41	10.30%	46.12%	9.15%	6.79%
r = 0.40	10.00%	46.11%	8.80%	6.65%

Table 6: Comparisons between different L1 loss under different sets.

MR \ Set L1	Reasonable	Heavy	Partial	Bare
Smooth	10.00%	46.11%	8.80%	6.65%
Vanilla	9.27%	46.34%	8.66%	5.62%

layer brings approximately 0.4% improvement with regard to MR. Table 3 shows no matter which solution we select, *SN* layer always contributes to performance improvement. Finally, as displayed in Table 4, we obtain our best performance under the help of *SN* layer. In conclusion, *SN* layer can substitute *BN* layer totally in our ACSP.

#### How important is the compressing width and vanilla L1 loss?

We talk about the contribution of the compressing width and vanilla L1 loss together in this part. Experiments show that, for Smooth L1, setting  $r$  in compressing width formula as 0.40 yields relatively good performance. And for vanilla L1 loss,  $r = 0.36$  is suitable. It should be mentioned that other settings may yield better results, but we choose to keep these settings in the following paragraphs(except where noted).

First, we only replace  $r = 0.41$  with  $r = 0.40$ , and the results are shown in Table 5. It can be seen that MR decreases about 0.3%.

Second, we compare the performance of Smooth L1 with vanilla L1 under respective optimized  $r$ . As displayed in Table 6, MR decreases to varying degrees on reasonable set, partial set, and bare set. However, MR increases about 0.2% on heavy set.

### 4.3 Comparison with the State of the Arts

We compare our ACSP with all existing state-of-the-art detectors(including preprint ones) on the validation set of CityPersons. The results are shown in Table 7. The evaluation metric is MR. To conduct a fair comparison, all methods are trained on the training set without any extra data(except ImageNet). When testing, the input scale is 1x. The top three results are highlighted in red, green and blue, respectively. Because the difference in training and test environment, *i.e.* most of other methods use Nvidia GTX 1080Ti GPU while we use Nvidia Tesla V100 GPU, time comparing is meaningless. As a result, it will not be reported in our table.

From the table, we can figure out that our ACSP achieves state-of-the-art on bare set and the second best performance on reasonable set, heavy set and

Table 7: Comparisons with state-of-the-arts on validation set: The evaluation metric is MR and the input scale is 1x. The top three results are highlighted in red, green and blue, respectively.

Method	Backbone	Reasonable	Heavy	Partial	Bare
FRCNN [49]	VGG-16	15.4%	-	-	-
FRCNN+Seg [49]	VGG-16	14.8%	-	-	-
TLL [37]	ResNet-50	15.5%	53.6%	17.2%	10.0%
TLL+MRF [37]	ResNet-50	14.4%	52.0%	15.9%	9.2%
Repulsion Loss [43]	ResNet-50	13.2%	56.9%	16.8%	7.6%
OR-CNN [52]	VGG-16	12.8%	55.7%	15.3%	6.7%
HBAN [25]	VGG-16	12.5%	48.1%	-	-
ALF [23]	ResNet-50	12.0%	51.9%	11.4%	8.4%
Adaptive NMS [21]	ResNet-50	11.9%	54.0%	11.4%	6.2%
CSP [24]	ResNet-50	11.0%	49.3%	10.4%	7.3%
MGAN [28]	VGG-16	10.5%	47.2%	-	-
PSC-Net [45]	VGG-16	10.4%	39.7%	-	-
APD [47]	ResNet-50	10.6%	49.8%	9.5%	7.1%
APD [47]	DLA-34	8.8%	46.6%	8.3%	5.8%
ACSP(Smooth L1)	ResNet-101	10.0%	46.1%	8.8%	6.7%
ACSP(Vanilla L1)	ResNet-101	9.3%	46.3%	8.7%	5.6%

partial set. On reasonable set, the best one, APD [47], uses more powerful backbone and other post process method. Without DLA-34, its MR will increase to 10.6% instead. On heavy set, without any special occlusion handling process, we outperform other special designed methods except for PSC-Net [45]. Also, we only lags behind APD [47] on partial set.

## 5 Conclusion

In this paper, we propose several improvements on original pedestrian detector CSP [24]. In this way, the training process of our ACSP is more robust. And we try to explain why we make these adaptations and why they make a difference. What's more, we propose a novel method to estimate the width of a bounding box. In addition, we explore some functions of Switchable Normalization which are not mentioned in its original paper [26]. Experiments are conducted on the CityPersons [49] and we achieve state-of-the-art on bare set and the second best performance on reasonable set, heavy set and partial set. In the future, it is interesting to explore the "representative

point" rather than the "center point" of pedestrian.

## 6 Acknowledgment

We thank Informatization Office of Beihang University for the supply of High Performance Computing Platform, which have 32 Nvidia Tesla V100 GPUs. This work is also supported by School of Mathematical Sciences, Beihang University.

## References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [2] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1844–1861, 2019.

- [3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [5] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. Fast feature pyramids for object detection. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1532–1545, 2014.
- [6] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–311. IEEE, 2009.
- [7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.
- [8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [11] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [12] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015.
- [13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [15] Tao Kong, Fuchun Sun, Huaping Liu, Yunling Jiang, and Jianbo Shi. Foveabox: Beyond anchor-based object detector. *arXiv preprint arXiv:1904.03797*, 2019.
- [16] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.
- [17] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.
- [18] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun. Detnet: Design backbone for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–350, 2018.
- [19] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *Proceedings. international conference on image processing*, volume 1, pages I–I. IEEE, 2002.

- [20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [21] Songtao Liu, Di Huang, and Yunhong 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.
- [22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multi-box detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [23] Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 618–634, 2018.
- [24] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan 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.
- [25] Ruiqi Lu and Huimin Ma. Semantic head enhanced pedestrian detection in a crowd. *arXiv preprint arXiv:1911.11985*, 2019.
- [26] Ping Luo, Jiamin Ren, Zhanglin Peng, Ruimao Zhang, and Jingyu Li. Differentiable learning-to-normalize via switchable normalization. *arXiv preprint arXiv:1806.10779*, 2018.
- [27] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. Local decorrelation for improved detection. *arXiv preprint arXiv:1406.1134*, 2014.
- [28] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4967–4975, 2019.
- [29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [31] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.
- [32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [33] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in neural information processing systems*, pages 901–909, 2016.
- [34] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, pages 2483–2493, 2018.
- [35] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowd-human: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.
- [36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

- [37] Tao Song, Leiyu Sun, Di Xie, Haiming Sun, and Shiliang Pu. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551, 2018.
- [38] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017.
- [39] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018.
- [40] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.
- [41] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.
- [42] Paul Viola, Michael Jones, et al. Robust real-time object detection. *International journal of computer vision*, 4(34-47):4, 2001.
- [43] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7774–7783, 2018.
- [44] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.
- [45] Jin Xie, Yanwei Pang, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Psc-net: Learning part spatial co-occurrence for occluded pedestrian detection. *arXiv preprint arXiv:2001.09252*, 2020.
- [46] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018.
- [47] Jialiang Zhang, Lixiang Lin, Yun-chen Chen, Yao Hu, Steven C. H. Hoi, and Jianke Zhu. CSID: center, scale, identity and density-aware pedestrian detection in a crowd. *CoRR*, abs/1910.09188, 2019.
- [48] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Towards reaching human performance in pedestrian detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):973–986, 2017.
- [49] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017.
- [50] Shanshan Zhang, Rodrigo Benenson, Bernt Schiele, et al. Filtered channel features for pedestrian detection. In *CVPR*, volume 1, page 4, 2015.
- [51] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6995–7003, 2018.
- [52] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–653, 2018.
- [53] Chunluan Zhou, Ming Yang, and Junsong Yuan. Discriminative feature transformation for occluded pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9557–9566, 2019.

- [54] Chunluan Zhou and Junsong Yuan. Bi-box regression for pedestrian detection and occlusion estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–151, 2018.
- [55] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.
- [56] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 850–859, 2019.
- [57] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 840–849, 2019.