# Box Guided Convolution for Pedestrian Detection

Jinpeng Li
Shengcai Liao*
Inception Institute of Artificial Intelligence (IIAI)
Masdar City, Abu Dhabi, UAE
jinpeng.li@inceptioniai.org
scliao@ieee.org

Hangzhi Jiang
School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
Institute of Automation, Chinese Academy of Sciences
Beijing, China
jianghangzhi2018@ia.ac.cn

Ling Shao
Inception Institute of Artificial Intelligence (IIAI)
Masdar City, Abu Dhabi, UAE
Mohamed bin Zayed University of Artificial Intelligence
Masdar City, Abu Dhabi, UAE
ling.shao@ieee.org

## ABSTRACT

Occlusions, scale variation and numerous false positives still represent fundamental challenges in pedestrian detection. Intuitively, different sizes of receptive fields and more attention to the visible parts are required for detecting pedestrians with various scales and occlusion levels, respectively. However, these challenges have not been addressed well by existing pedestrian detectors. This paper presents a novel convolutional network, denoted as box guided convolution network (BGCNet), to tackle these challenges simultaneously in a unified framework. In particular, we proposed a box guided convolution (BGC) that can dynamically adjust the sizes of convolution kernels guided by the predicted bounding boxes. In this way, BGCNet provides position-aware receptive fields to address the challenge of large variations of scales. In addition, for the issue of heavy occlusion, the kernel parameters of BGC are spatially localized around the salient and mostly visible key points of a pedestrian, such as the head and foot, to effectively capture high-level semantic features to help detection. Furthermore, a local maximum (LM) loss is introduced to depress false positives and highlight true positives by forcing positives, rather than negatives, as local maximums, without any additional inference burden. We evaluate BGCNet on popular pedestrian detection benchmarks, and achieve the state-of-the-art results, with the significant performance improvement on heavily occluded and small-scale pedestrians.

## CCS CONCEPTS

• **Computing methodologies** → **Object detection**; *Vision for robotics*.

## KEYWORDS

Pedestrian Detection; Box Guided Convolution; Receptive Fields; Scale Variation

*Shengcai Liao is the Corresponding Author.


*MM '20, October 12–16, 2020, Seattle, WA, USA*

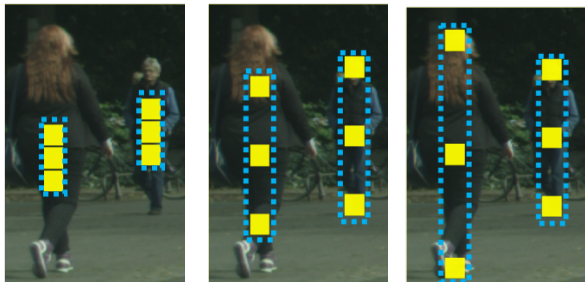
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413989

**ACM Reference Format:**
Jinpeng Li, Shengcai Liao, Hangzhi Jiang, and Ling Shao. 2020. Box Guided Convolution for Pedestrian Detection. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413989

## 1 INTRODUCTION

Pedestrian detection serves as a key and challenging task in multimedia community, due to its wide range of applications, such as video surveillance, intelligent vehicles and robotics. In recent years, the performance of pedestrian detection has been greatly driven by the great progress of convolutional neural networks (CNNs). However, it is worth noting that the pedestrian detection still suffers from the fundamental challenges of occlusions, scale variation and numerous false positives.

As is well known, various scales of pedestrians are inherent for images, due to the principle of perspective projection which results in larger sizes of near objects and smaller sizes of distant objects. Although current pedestrian detectors perform well on the evaluation indicator of average log miss rates ($MR^{-2}$), the detection performance is still sensitive to the pedestrian scales. Considering this challenge, one-stage methods put efforts in designing more dedicated architectures to improve the performance on various scales situations. MS-CNN [1] uses pyramid feature maps to generate multi-level predictions. ALFNet [28] proposes an asymptotic localization method to gradually refine the default anchors for improving detection accuracy. However, the receptive fields of these methods are fixed and cannot be dynamically adjusted according to the real scales of pedestrians. For multi-stage methods [3, 36], RoI pooling [36] is leveraged to reshape region proposals into a fixed size, and then fully connected layers are applied to force the receptive fields to cover the entirety of objects with various scales. Nevertheless, the computation of proposals in the second stage are not shared, and the rectangular proposals do not consider the inherent appearance of pedestrians, which introduces extensive background information and further confuses the classifier, as pointed out by [38]. To more efficiently handle the difficulty of various scales and focus on the physical characteristic of pedestrians, the proposed BGC layer dynamically adjusts the kernel sizes in a position and scale sensitive manner and places the kernel parameters around the body parts. In comparison to RoI pooling, the input and output feature maps of BGC layer have the same dimension, thus it can be used as the normal convolutional layer and



(a) Convolution (b) Dilated Convolution (c) BGC

**Figure 1: Comparisons of convolution, dilated convolution and our proposed BGC with the kernel size of $3\times1$. (a) Convolution has the fixed receptive field. (b) Dilated convolution can control the receptive field by the hyper parameter of dilation rate, but the receptive fields are same among different positions. (c) BGC can adaptively adjust the sizes of kernel under the guidance of predicted bounding boxes, thus having the position-aware receptive fields.**

easily integrated into most CNN based detectors. Besides, due to the computations after the RoI pooling layer are not shared, heuristic methods and NMS are greatly relied on to abandon the proposals with low confidences, which increases the training difficulties and undermines the accuracy. Instead of individually processing each proposal, the computations of features in overlapping areas are consistently shared across all layers in our BGCNet, which doesn't need any heuristics and keeps the full information.

On the other hand, the detection performance of the pedestrian detectors on heavy occluded pedestrians is still much worse than the results on non-occluded and partially occluded pedestrians. Recently, more and more works have been devoted to tackle this challenge. Most of these methods exploit the information obtained from visible pedestrian parts. For instance, [32, 42, 53] separately learn part detectors and heuristically integrate detection results together. Bi-box [54] proposes a Faster R-CNN based network with two branches to simultaneously detect the full bounding box and visible part bounding box. However, multiple detectors or branches for detecting extra boxes increase the computational cost in both the training and testing phases. Besides, additional annotations are usually required, which impedes their practical application.

Even for the state-of-the-art pedestrian detectors, detecting pedestrians with various scales and occlusions is still very challenging. For example, CSP [29] is an anchor-free one-stage pedestrian detector which achieves the state-of-the-art results on CityPersons [50] and Caltech [13] datasets. However, it performs four times worse on small scale pedestrians than medium ones, and performs five times worse on heavy occluded pedestrians than non-occluded ones. The main reasons of these performance gaps include: (1) the fixed sizes of receptive fields in the one-stage structure are hard to detect various scales pedestrians, (2) the features focused on center points are not robust enough to handle occlusion, and (3) the false positive results are hard to remove in crowd scenes. To tackle these challenges, we propose a new pedestrian detector called BGCNet, which benefits from the novel BGC layer, LM loss and consistent feature representation. Specifically, the BGC layer has two advantages: (1) The receptive fields are position-aware and scale adaptive, which can be dynamically adjusted under the guidance of predicted bounding boxes for effectively detecting pedestrians of various scales. (2) The kernel parameters of BGC are placed around the mostly visible and semantic key parts around pedestrians, such as the head and foot, which improves the robustness on detecting occluded situations. On the other hand, in order to decrease the false positives in crowd, the proposed LM loss compares the scores within neighbouring regions, encourages the positive points to be the local maximums, and suppresses the negative points to be smaller than the local maximum. In addition, we employ the consistent and multi-scale feature maps in the backbone [40] to help detect occluded and small scale pedestrians.

In summary, this paper provides three main contributions to pedestrian detection studies:

- We propose a box guided convolution which can dynamically adjust the sizes of kernels under the guidance of bounding boxes and focus on the hard occluded parts to meet the challenges of various scales and occlusions in pedestrian detection.
- We design a novel local maximum loss to constrain the relative magnitude of the scores within the neighbouring region for reducing the false positive results.
- We achieve state-of-the-art results on the CityPersons [50] and Caltech [13] datasets, with significant improvement in the cases of various scales and occlusion levels.

## 2 RELATED WORK

**Generic object detection.** Recent advances in generic object detection have been driven by the development of CNN. Fast R-CNN [17] inherits the classical sliding-window paradigm to formulate a deep learning based two-stage detection framework. Specifically, the first stage employs a hand-crafted proposal generator [44], and the second stage uses a CNN to classify these regions and regress more accurate bounding boxes. Faster R-CNN [36] extends the traditional proposal generation methods to CNN based region proposal network (RPN). Cascade R-CNN [3] uses the cascade architecture to iteratively regress the bounding boxes and refine the detection performance in a multi-stage manner. One-stage methods [5, 25, 27, 35] convert the RPN into multi-class classification and remove the R-CNN sub-networks, which accelerates the detection. To narrow the performance gap between one-stage methods and multi-stage methods, RefineDet [52] and HSD [4] use the cascade structure to refine the results. Recently, anchor-free detectors [24, 55, 56] have attracted more attention because of their high performance and simplicity compared to the anchor-based methods. Instead of classifying anchors, anchor-free methods directly classify the centers [56], corners [24] or extreme points [55] of objects. This not only avoids the need for carefully choosing the hyper-parameters of anchors, but also unifies the object detection and semantic classification into similar dense prediction frameworks.

**Pedestrian detection.** Since the pioneering detection framework of Viola and Jones [45], traditional pedestrian detection methods [6, 15, 22, 34] use the sliding-window strategy to extract hand-crafted features, such as HOG [10] and LBP [21], and feed them to the classifier to identify each pedestrian. Inspired by the great success of deep learning in generic object detection, early works[2,

23, 42] employ ACF [12] or LDCF [31] detectors to generate proposals, and leverage CNNs to classify them. Instead of combining traditional detectors with CNN, Bi-box [54] and Occlusion-Aware R-CNN [51] leverage the vanilla Faster R-CNN with appropriate modifications and achieve improved results. Although deep learning based methods significantly improve the detection performance, various scales and heavy occluded situations are still challenging.

For detecting pedestrians with various scales, MS-CNN [1] employs pyramid features from the network backbone to match different scale objects. In [14], MS-CNN is further combined with bi-directional LSTM to exploit the nonlinear relationships among multi-scale features. Inspired by the line annotation in CityPersons and Caltech, TLL [39] locates the more scale insensitive topological elements of top and bottom vertices and link lines for detecting multi-scale pedestrians. However, it is not end-to-end trainable due to the MRF based post-processing manner.

In addition, various methods have put efforts towards improving the detection performance on occluded pedestrians. In [30, 32, 42, 53], manually designed or learned features are employed to train detectors for multiple parts to handle different occlusion patterns. However, these methods have to heuristically ensemble the detectors, which increases the computational costs. Adaptive-NMS [26] learns the density scores and uses a dynamic bounding boxes suppression strategy to handle the crowd and occlusion cases. Repulsion loss is proposed in [46] to keep the proposals away from surrounding objects for the crowd scenes.

However, most of the above methods exclusively focus on various scale or occlusion cases. In contrast, the proposed BGCNet aims to simultaneously solve these fundamental challenges of scale variation and occlusion by the simple and effective architecture.

**Position-aware receptive fields.** Multi-stage detectors [36] [3] obtain the ability of position-aware receptive fields from the RoI pooling layer which crops and resizes the features in region proposals to the same size. However, the computation of second stage is not shared among proposals, which makes the two-stage detectors more time-consuming. The deformable convolution [9] uses an extra layer to learn the offsets of kernel for free transforming the shape and size of the receptive field on each position. The scale-adaptive convolution [47, 48] is proposed for scene parsing and medical image segmentation, which regresses the position-aware scale coefficients to resize the convolution patches for adjusting the receptive fields. However, the offsets and scale coefficients are implicitly learned without any direct supervision, which could change the topology of kernel and may shift the kernel from the focusing object to other objects. In contrast, the kernel of proposed BGC is adjusted under the guidance of the predicted bounding box, which does not need to learn additional parameters and assures that the size of receptive field at each position is same as the extent of target pedestrian.

# 3 OUR APPROACH

## 3.1 Architecture

**Overall architecture.** To keep the model simple and efficient, the proposed BGCNet adopts the one-stage anchor-free detector, and its overall architecture is illustrated in the Figure 2. Following the common designs of anchor-free detectors [29, 43], BGCNet can be summarized as three components, including: (1) a backbone for generating feature maps with multi-scale resolutions, (2) a neck for unifying the sizes of feature maps, and (3) a detection head for generating the detection results. Specifically, the network backbone processes the input image and produces the pyramid feature maps. Then the network neck upsamples these feature maps to 1/4 of the input sizes through deconvolutional layers and concatenates them together as inputs for the detection head. The detection head contains three branches which are used to regress heights, regress center offsets and classify the centers of pedestrians in a pixel-wise prediction manner, respectively. To handle the challenges of various scales and occlusions, the proposed BGCNet embeds the BGC layer and LM loss into the classification branch. The details of these components are described in Sections **3.2** and **3.3**.

The proposed BGCNet is optimized by the multi-task loss function as follows:

$$L = \lambda_h L_{height} + \lambda_o L_{offset} + \lambda_c L_{cls} + \lambda_l L_{local} \tag{1}$$

where $L_{height}$ and $L_{offset}$ are smooth L1 losses for height and offset regression, $L_{cls}$ is the focal loss [25] for center classification, $L_{local}$ is our proposed LM loss to reduce false positives, $\lambda_h$, $\lambda_o$, $\lambda_c$, and $\lambda_l$ are the weights for height regression, offset regression, center classification and LM loss, respectively.

## 3.2 Box Guided Convolution

Scale variation is the fundamental challenge in pedestrian detection, which is hard to handle for the one-stage detectors with fixed receptive fields. To solve this challenge, a box guided convolution (BGC) layer, which can adaptively adjust the dilation rates guided by the predicted bounding boxes, is proposed to provide one-stage methods the ability of position-aware receptive fields. Firstly, we review the dilated convolutional layer [7]. Let the output $y$ of dilated convolution at position $(i, j)$ be defined as follows:

$$y[i,j] = \sum_{k_y=-n}^{n} \sum_{k_x=-m}^{m} x[i + r \cdot k_y, j + r \cdot k_x] \cdot w[k_y, k_x] \tag{2}$$

where $x$ is the input feature map, $w$ is the convolution kernel, the kernel size is $[2n+1, 2m+1]$, $r$ is the dilation rate, and the dilated kernel size is $[2rn+1, 2rm+1]$. Since $r$ is shared across all positions, the receptive field is same for all the pixels in a feature map. So, as illustrated in Figure 1(b), the kernels in dilated convolution cannot capture the entire body of pedestrians with different scales. In contrast, the proposed BGC layer uses the adaptive and position-aware dilation rate $r^{ij} = (r_y^{ij}, r_x^{ij})$ to allow different pixels within a feature map to have different dilated kernel sizes. The BGC is defined as follows:

$$y[i,j] = \sum_{k_y=-n}^{n} \sum_{k_x=-m}^{m} x[i + r_y^{ij} \cdot k_y + o_y^{ij}, j + r_x^{ij} \cdot k_x + o_x^{ij}] * w[k_y, k_x] \tag{3}$$

where $[2r_y^{ij}n+1, 2r_x^{ij}m+1]$ is the dilated kernel size, and $(o_y^{ij}, o_x^{ij})$ is the offset of the convolution kernel.

In the height and offset regression branches, BGCNet produces the bounding box $B_{ij} = (h_{i,j}, w_{ij}, i + \hat{o}_y^{ij}, j + \hat{o}_x^{ij})$ at each position

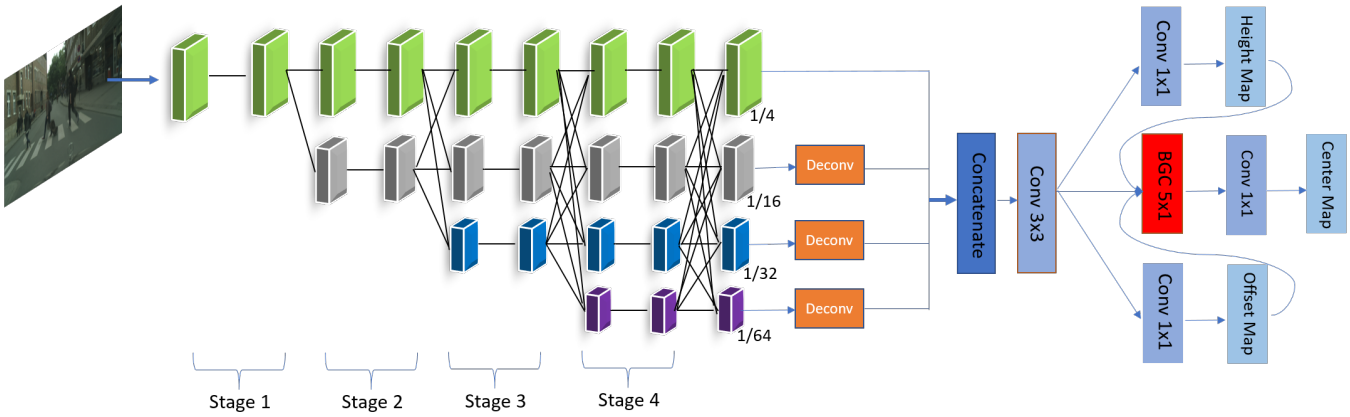



**Figure 2: Overall architecture of BGCNet. In the backbone, feature maps with the same color represent a stream and have the same resolution. There are four parallel streams in BGCNet for generating the pyramid feature maps.**

$(i, j)$, where $h_{ij}$, $w_{ij}$ denote the height and width of the box, respectively, and $(i + \hat{o}_y^{ij}, j + \hat{o}_x^{ij})$ is the coordinate of box center. To capture the full information of pedestrians with different scales, the BGC kernel is adjusted to the same size as the predicted bounding box $B_{ij}$, such that:

$$2r_y^{ij} n + 1 = h_{ij} \tag{4}$$

$$2r_x^{ij} m + 1 = w_{ij} \tag{5}$$

$$o_y^{ij} = \hat{o}_y^{ij} \tag{6}$$

$$o_x^{ij} = \hat{o}_x^{ij} \tag{7}$$

so the dilated rate in BGC is calculated as follow:

$$r_y^{ij} = \begin{cases} (h_{ij} - 1)/(2n), & \text{if } n \neq 1. \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

$$r_x^{ij} = \begin{cases} (w_{ij} - 1)/(2m), & \text{if } m \neq 1. \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

As illustrated in Figure 1(c), BGC even places the kernel parameters within the detected pedestrians. Because $(r_y^{ij}, r_x^{ij})$ are basically fractional, the value of $x[i+r_y^{ij} \cdot k_y + o_y^{ij}, j+r_x^{ij} \cdot k_x + o_x^{ij}]$ is produced by bilinear interpolation, which is inspired by [9].

Besides, BGC also makes full use of the physical features of pedestrians to handle the challenge of heavy occlusion in pedestrian detection. Specifically, as illustrated in Figure 1(c), our BGC layer directly places the kernel parameters on the center, top-center and bottom-center vertices of bounding box. These key points are closer to the body parts of head and foot, which contain more obvious image features and semantic information. Because they are mostly visible in the occluded pedestrians, the features learned by BGC are more distinguishable and robust, which helps to increase the detection performance in crowd scene. The methods designed for occluded pedestrian detection usually heavily rely on the extra annotations of visible bounding boxes or head boxes. In contrast, our BGC obtains the features suitable for occluded pedestrians solely from the full bounding boxes, which is more elegant and low-cost. TLL [39] also tries to leverage the key points from full bounding box. It designs a bottom-up method to detect the top and down vertices of pedestrians and then group them for improving the detection performance on heavy occluded and small scale pedestrians. However, the post-processing step of grouping increases the computational cost and may incur mismatches. In contrast to the bottom-up design in TLL, our BGC layer is a top-down method which leverages the detected bounding box to locate the key points and capture their features. Our top-down design avoids additional matching or post-processing steps and can utilize more key points than TLL. Furthermore, the BGC kernel is designed to be asymmetric for the datasets with line annotations, with the size of $[2r_y^{ij} n + 1, 1]$. The scale kernel focuses more on the vertical axis features than the horizontal ones, which not only improves the robustness but also reduces the number of parameters.

### 3.3 Local Maximum Loss

Non-maximum suppression (NMS) is necessary in CNN based pedestrian detectors, which aims at retaining one positive point with the highest score per object while suppressing all the other points. However, the repeated convolution operations are prone to produce blur score maps where the negative points around positive ones may also have enough high scores that NMS can not suppress. This phenomenon is especially serious in the crowded scene, which results many false positive boxes among crowded pedestrians.

The classification losses, such as cross entropy loss and focal loss, only individually supervise each predicted score and ignore the relative magnitude among neighbouring points, which limits their abilities to tackle this challenge. Therefore, we propose a local maximum (LM) loss to constrain the relative magnitude of the pixels scores within a local region, in order to reduce the false positives and increase the quality of true positives. Specifically, the LM loss contains two components as follows:

$$L_{local} = L_{pos} + \alpha * L_{neg} \tag{10}$$

where $L_{pos}$ is the positive term which highlights the positive points to encourage them to be the local maximums, while $L_{neg}$ is the negative term which suppresses the negative points to prevent them from being the local maximum, and $\alpha$ is the coefficient to balance the positive and negative terms. For simplicity, the proposed LM loss only considers the relationship between the scores of points, and does not compare the IOUs of different boxes, as NMS does. It is

**Table 1: Ablation experiments of BGCNet on the CityPersons [50] benchmark. LM is the local maximum loss. CF is the consistent feature representation. The results of all detectors are tested on the original image size. The evaluation metric is $MR^{-2}$.**

| Method | Backbone | BGC | LM | CF | Reasonable | Occlusion Level | | | Scale Level | | | Test Time |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Heavy | Partial | Bare | Small | Medium | Large | |
| Baseline | ResNet-50 | | | | 10.90 | 49.56 | 10.42 | 7.32 | 14.76 | 4.92 | 6.96 | 196ms/img |
| BGCNet | ResNet-50 | | ✓ | | 10.45 | 46.59 | 9.90 | 6.58 | 14.12 | 3.68 | 6.73 | 196ms/img |
| BGCNet | ResNet-50 | ✓ | | | 9.84 | 45.80 | 9.16 | 6.49 | 14.43 | 3.89 | 5.59 | 200ms/img |
| BGCNet | ResNet-50 | ✓ | ✓ | | 9.39 | 45.94 | 9.01 | 6.43 | 14.07 | 3.69 | 5.50 | 200ms/img |
| BGCNet | HRNet-w32 | ✓ | ✓ | ✓ | **8.83** | **43.91** | **7.95** | **6.14** | **11.55** | **2.61** | **5.30** | 156ms/img |

parallel to the original classification loss and directly supervises the classification map, without extra burden for the network inference.

**Positive term.** In the classification branch of BGCNet, only the center points of pedestrians are positives, and all the other points are negatives. The positive term forces the positive point to be larger than all the surrounding negative points, which is equivalent to being larger than the local maximum of surrounding negatives. Given the position $(i, j)$, the local maximum value within a neighbouring area of $(i, j)$ is defined as:

$$m(i, j, r) = \max_{\substack{-r \le y \le r \wedge y \neq 0 \\ -r \le x \le r \wedge x \neq 0}} (s_{i+y, j+x}) \tag{11}$$

where r is the neighbouring radius, and $s_{ij}$ is the classification score at position $(i, j)$.

The positive term is designed as follows:

$$L_{pos} = \frac{\sum_{(i,j) \in P_+} \ell_{pos}(s(i, j), m(i, j, r))}{|P_+|} \tag{12}$$

$$\ell_{pos}(a, b) = max(b + \delta - a, 0) \tag{13}$$

where $P_+$ is the set of all positive points, $|P_+|$ is the number of points in $P_+$, and $\delta$ is the margin. $\ell_{pos}$ pushes the $s(i, j)$ to be bigger than $m(i, j, r)$ until the difference is beyond $\delta$. In this way, $L_{pos}$ can reduce the possibility of the positive point being suppressed by its surrounding points in NMS.

**Negative term.** When a negative point is the local maximum, it is likely to be retained by the post-processing step and becomes the false positive. Therefore, the negative term of the LM loss forces the scores of negative points to be smaller than the local maximum of their neighbours. The definition of negative term $L_{neg}$ is as follows:

$$L_{neg} = \frac{\sum_{(i,j) \in P_-} \ell_{neg}(s(i, j), m(i, j, r))}{|P_-|} \tag{14}$$

$$\ell_{neg}(a, b) = max(a - b, 0) \tag{15}$$

where $P_-$ is the set of all negative points within the ground truth bounding boxes. Compared with the positive term, we omit the margin in $\ell_{neg}$, because the score gap between different negative points should not be large. In Eq. (14) and Eq. (15), only the negative points with scores higher than their neighbours are suppressed, which smooths the score maps of negative points, so as to reduce the number of false positives.

### 3.4 Consistent Feature Representation

The backbones in generic object and pedestrian detectors usually have multiple stages to gradually downsample the feature map and extract different level features. However, for detecting various scales pedestrians, detectors [29, 43] force each stage to simultaneously learn the feature representations of different levels, which increases the learning difficulty. For example, the feature maps in the second stage simultaneously have to represent the low-level features used as the input of the third stage, and represent the high-level semantic features used for small-scale pedestrian detection. Inspired by [40], the proposed BGCNet employs the consistent feature representation in the backbone to generate multi-scale feature maps with the similar semantic levels to resolve feature conflict. As shown in Figure 2, our backbone is decomposed to two dimensions including the streams and stages. Each stream contains multiple convolutional layers with the same resolutions of all feature maps constant across different stages. In each stage, the feature maps from different streams have different resolutions to provide pyramid features. In the end of each stage, the feature maps of a stream are resized and added to all the other streams, in order to exchange and fuse the information. We extract pyramid feature maps from the last stage of backbone, which have the similar high-level semantic features. Compared to the detectors using one-stream backbones, BGCNet only uses the feature maps from a stage as the input of the next stage, without the conflict of learning multiple levels of representation. Since the better feature representation, BGCNet can improve the detection performance under the lower computational cost.

## 4 EXPERIMENTS

In this section, we first introduce the datasets and evaluation metrics for pedestrian detection, followed by the implementation details of the proposed BGCNet. Then, we conduct ablation experiments to demonstrate the effectiveness of different components in BGCNet. Finally, we compare the performance of BGCNet with other state-of-the-art pedestrian detectors.

### 4.1 Datasets and Evaluation Metrics

We evaluate our proposed BGCNet on the public CityPersons [50] and Caltech [13] datasets which contain a large number of pedestrians with various scales and different levels of occlusion.

**CityPersons.** The CityPersons dataset is a challenging pedestrian detection benchmark proposed in [50]. It is built on top of the CityScapes [8] dataset which is recorded from the street views of multiple European cities and has pixel-wise semantic segmentation annotations. Following the state-of-art pedestrian detectors, we use the official training set with 2,975 images for training and the validation set with 500 images for evaluating.

**Table 2: Comparisons of BGC layers with different kernel shapes. ResNet-50 is used as the backbone. The evaluation metric is $MR^{-2}$.**

| Kernel Shape | Height | Width | Reasonable | Scale Level | | |
|---|---|---|---|---|---|---|
| | | | | Small | Medium | Large |
| square kernel | 3 | 3 | 10.49 | 15.06 | **3.48** | 6.46 |
| horizontal kernel | 1 | 3 | 11.11 | 15.79 | 3.96 | 6.85 |
| vertical kernel | 3 | 1 | **10.34** | **13.94** | 4.10 | **6.07** |

**Caltech.** The Caltech dataset is a popular pedestrian detection benchmark, which is composed of 10 hours of video recorded in Los Angeles, USA. The video frames in the Caltech are sampled and divided into training set and testing set with 42,500 images and 4,024 images, respectively. The ground truths are further refined through the line annotation method by Zhang *et al.* [49]. We use these new annotations for both the training and the testing phases.

**Evaluation Metrics.** The standard log average Miss Rate over False Positive Per Image (FPPI) range of $[10^{-2}, 10^{0}]$ ($MR^{-2}$) is used as the evaluation metric for all the experiments, and a lower $MR^{-2}$ indicates better results. Since various scales and occlusion are the most challenging problems in pedestrian detection, our models are tested across different occlusion levels (reasonable, bare, partial and heavy) and different scales (small, medium and large).

## 4.2 Implementation Details

The proposed BGCNet is implemented under the framework of Pytorch [33]. The backbones of both ResNet-50 [20] and HRNet-w32 [40] are pre-trained on the ImageNet [11] dataset. The other parameters are randomly initialized by the MSRA method [19]. We use the Adam method to optimize the models, and the warm-up strategy [18] to adjust the learning rate. For a fair comparison with other methods, we apply the same data augmentation methods and the training stabilization method of mean teacher [41] as the state-of-the-art method CSP [29]. The kernel size of BGC layer is $5 \times 1$, unless other specification. The loss weights of $\lambda_h$, $\lambda_o$, $\lambda_c$, and $\lambda_l$ are set to 1, 0.1, 0.01 and 0.1, respectively. The neighbouring radius and coefficient $\alpha$ in LM loss are 1 and 0.1, respectively. For the CityPersons dataset, the models are trained on four Tesla V100 GPUs with a mini-batch of four images per GPU. The training images are crop to a resolution of $640 \times 1280$ and the models are trained for 37k iterations. The initial learning rate is $2 \times 10^{-4}$, which is decreased to $2 \times 10^{-5}$ and $2 \times 10^{-6}$ after 20k iterations and 30k iterations, respectively. For the Caltech dataset, we train the network on one GPU with a mini-batch of 16 images. We crop the training images into $336 \times 448$. The models are trained for 25k iterations, with an initial learning rate of $2 \times 10^{-4}$, which is decreased to $2 \times 10^{-5}$ and $2 \times 10^{-6}$ after 14k and 20k iterations.

## 4.3 Ablation Analysis on BGCNet components

We conduct the ablation study on the validation set of CityPersons [50] to demonstrate the effectiveness of BGCNet components. The design of BGC layer and LM loss is general and could be easily integrated to any one-stage detectors. To prove that the challenges they tackle are common even for the detectors with high performance, the state-of-the-art pedestrian detector of CSP is used as the baseline. In Figure 3, we compare the results of the baseline CSP detector and BGCNet. The implemented baseline achieves the performance of 10.90% $MR^{-2}$ on the reasonable setting, which is already better than most of the state-of-the-art methods. We compare multiple ablated versions of BGCNet against the baseline. Components under comparison include the BGC layer, the LM loss and the consistent feature representation. The results are shown in Table 1.

**Table 3: Comparisons of BGC layers with different kernel heights. ResNet-50 is used as the backbone. The evaluation metric is $MR^{-2}$.**

| Kernel Height | Reasonable | Scale Level | | |
|---|---|---|---|---|
| | | Small | Medium | Large |
| 3 | 10.34 | **13.94** | 4.10 | 6.07 |
| 5 | **9.84** | 14.43 | **3.89** | **5.59** |
| 7 | 10.45 | 14.68 | 4.12 | 6.23 |

**Table 4: Comparisons of BGCNet with different types of convolutional layer in the classification branch. ResNet-50 is used as the backbone. The evaluation metric is $MR^{-2}$.**

| Convolution Types | dilation rate | Reasonable | Scale Level | | |
|---|---|---|---|---|---|
| | | | Small | Medium | Large |
| Dilated Convolution | 1 | 10.85 | 14.76 | 4.63 | 6.21 |
| Dilated Convolution | 2 | 10.77 | 14.08 | 3.89 | 6.75 |
| Dilated Convolution | 4 | 10.39 | **13.16** | 4.26 | 5.91 |
| Dilated Convolution | 8 | 10.55 | 15.65 | 4.09 | 5.89 |
| Deformable Convolution | None | 11.07 | 15.22 | **3.71** | 6.63 |
| BGC (ours) | Adaptive | **9.84** | 14.43 | 3.89 | **5.59** |

**BGC Layer.** Firstly, by fixing the ResNet-50 backbone, when the BGC based classification branch is applied to the baseline, the performance is 9.84% $MR^{-2}$ in the reasonable subset, with the significant improvement of 1.06% $MR^{-2}$. In addition, the results on all the other occlusion and scale levels also outperform the baseline, which fully demonstrates the effectiveness of the proposed BGC in handling the problem of various pedestrian scales and occlusion.

**LM Loss.** When only the LM loss is added to the baseline, the $MR^{-2}$s of the detector are improved from 10.90% and 48.68% to 10.45% and 46.59% on the reasonable and heavy occlusion subsets, respectively. This demonstrates the effectiveness of the LM loss in highlighting the true positives and suppressing the false positives. Furthermore, combing in the BGC and LM loss together further improves the performance on the reasonable and heavy subsets to 9.39% $MR^{-2}$ and 45.94% $MR^{-2}$, respectively.

**Consistent Feature Representation.** Finally, the backbone of BGCNet is replaced with HRNet-w32 in order to verify the effectiveness of the consistent feature representation. As shown in the last two rows of Table 1, BGCNet with consistent feature representation significantly improves the results from 9.39% $MR^{-2}$ to 8.83% $MR^{-2}$ on the reasonable subset, and achieves the best performance on all scale and occlusion levels.

Next, we conduct an ablation study for the design of BGC layer on the CityPersons validation set.

**Asymmetric Kernel.** The proposed BGC leverages the vertical kernel to capture the pedestrian information along the top to bottom topological line. To prove that the features obtained by the vertical kernels are more discriminative, we compare the detection performance among the adaptive convolution kernels with three different settings, i.e. square kernel $3 \times 3$, horizontal kernel $1 \times 3$ and vertical kernel $3 \times 1$. The results are summarized in Table 2. As can be seen, the BGC with the vertical kernel achieves

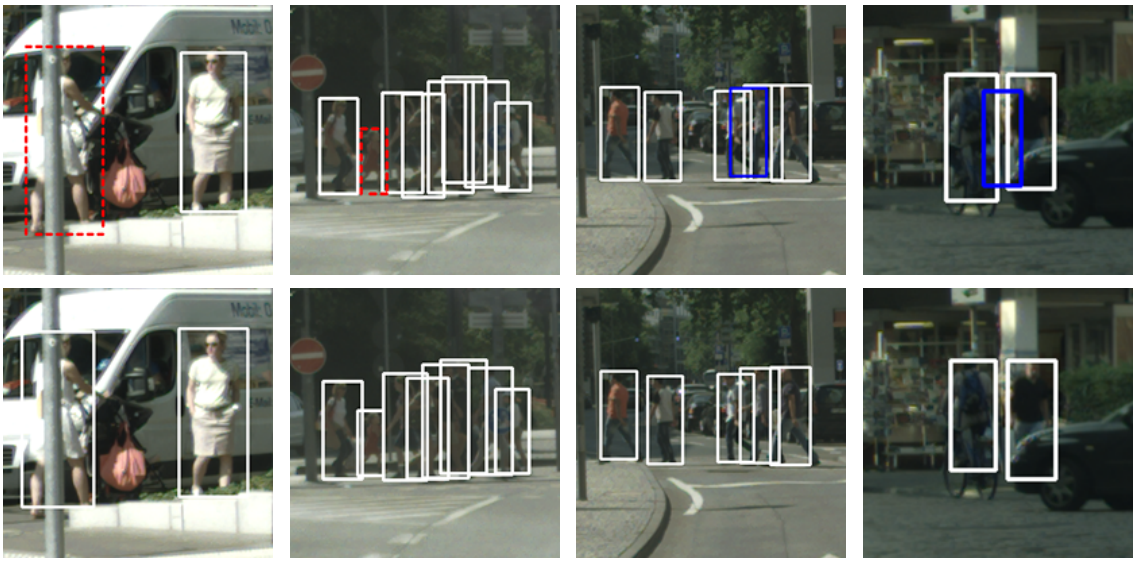


**Figure 3: Qualitative comparisons of the baseline CSP detector and the proposed BGCNet. First row is CSP. Second row is BGCNet. The white, red and blue boxes represent the true positives, false negatives, and false positives, respectively.**

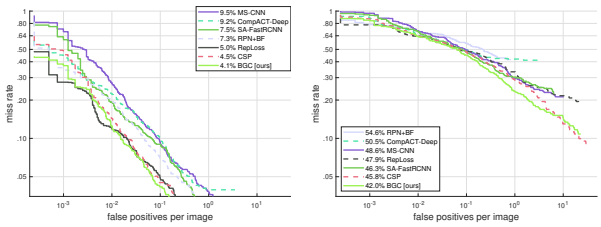
the best detection performance of 10.34% $MR^{-2}$ on the reasonable subset and also outperforms the square and horizontal kernels on all the scale levels. Although the widely used square kernel has three times the parameters than the vertical kernel, it still performs worse than the vertical kernel. In addition, the horizontal kernel achieves the lowest accuracy, even lower than the baseline. This is probably because the key points from top and down sides are around the head and the foot, which have more salient and semantic features, but the points from left and right sides do not have clearly semantic features, especially when the pedestrian width is automatically generated based on the fixed aspect ratio in the line annotated pedestrian datasets. Note that these experiments are only conducted on pedestrian detection, and the square BGC may be more suitable for generic object detection benchmarks using box annotations. What we want to show here is that BGC with the vertical kernel is more effective and efficient for extracting features from pedestrians with line annotation.

**Kernel Height.** The kernel height in BGC determines the number of feature points vertically sampled from the predicted bounding box. Figure 1(c) shows that the $3 \times 1$ kernel places the parameters on the top, center and down topological vertices of pedestrian, in contrast, the $5 \times 1$ kernel has two more parameters, thus capturing more internal information. We conduct a comparison among three different kernel sizes including $3 \times 1$, $5 \times 1$ and $7 \times 1$. The results are shown in Table 3. Compared to the $3 \times 1$, $7 \times 1$ kernels, the BGC with $5 \times 1$ kernel achieves the best detection performance of 9.84% $MR^{-2}$, 3.89% $MR^{-2}$ and 5.59% $MR^{-2}$ on the reasonable, medium and large scale subsets, respectively. Only the accuracy on the small scale subsets is worse than the result of the $3 \times 1$ kernel. The reason behind the performance difference among different kernel sizes may be that the small kernel cannot capture enough features inside the predicted box, while the large kernel size may cause overfitting.

**Box Guided Dilation.** To demonstrate the effectiveness of the box guided dilation in BGC layer, we compare the BGC with the dilated convolution [7] and the deformable convolution [9] which are widely used in the detection tasks. Unlike the proposed BGC, the dilation rate of the dilated convolution is a hyper parameter, which cannot be dynamically adjusted. Instead of using the dilation manner, the deformable convolution regresses the offset of each kernel parameter by a $1 \times 1$ convolutional layer. We construct five comparable networks, by replacing the BGC layer with the dilated convolution with the dilation rate of 1, 2, 4, 8, and the deformable convolution, respectively. The dilated convolution with the dilation rate of 1 is the same as the original convolution. The kernel sizes of all convolutions are set to $5 \times 1$. As shown in Table 4, the BGC layer achieves the best performance on the reasonable subset, and outperforms the dilated convolution with a dilation rate of 4, which is the second best, by 0.55% $MR^{-2}$ on the reasonable subsets. Although BGC does not always achieve the best result in each scale level, the higher performance on reasonable set and more balanced performance on various scale levels show that BGC layer can improve the overall performance instead of being specifically optimized for a single subset. Among the networks with dilated convolutions, the detection performance on the reasonable subset increases with dilation rate until the value of 4. The inference time of baseline with dilation rate of 4 is 199ms, which is similar to BGCNet. The dilation rate of 8 is 0.16% $MR^{-2}$ worse than the dilation rate of 4 on the reasonable subset. This is probably because the fixed dilation rate struggles to handle various scales of pedestrians. In addition, the deformable convolution is worse than the dilation convolution and the BGC. Our explanation is that the offsets in the deformable convolution are learned without any direct supervision and thus have difficulties in capturing the full information of pedestrians. Moreover, the deformable convolution needs an extra convolutional layer to regress the offsets of each kernel parameter and does not constrain the shape and size of final kernel. In contrast, BGC directly calculates the dilation rate based on the predicted bounding box without relying any other layers and keeps the relative spatial

**Table 5: Comparison with the state-of-the-arts on the CityPersons validation set [50]. Scale represents the upsampling level of the original image. Red and blue numbers indicate the best and second performance, respectively.**

| Method | Scale | Backbone | Reasonable | Occlusion Level | | | Scale Level | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Heavy | Partial | Bare | Small | Medium | Large |
| Adapted Faster R-CNN [50] | ×1.3 | VGG16 | 12.8 | - | - | - | - | - | - |
| TLL [39] | ×1 | ResNet-50 | 15.5 | 53.6 | 17.2 | 10.0 | - | - | - |
| TLL+MRF [39] | ×1 | ResNet-50 | 14.4 | 52.0 | 15.9 | 9.2 | - | - | - |
| Repulsion Loss [46] | ×1.3 | ResNet-50 | 11.6 | 55.3 | 14.8 | 7.0 | - | - | - |
| Bi-box [54] | ×1.3 | VGG16 | 11.2 | - | - | - | - | - | - |
| OR-CNN [51] | ×1.3 | VGG16 | 11.0 | 51.3 | 13.7 | **5.9** | - | - | - |
| ALFNet [28] | ×1 | ResNet-50 | 12.0 | 51.9 | 11.4 | 8.4 | 19.0 | 5.7 | 6.6 |
| CSP [29] | ×1 | ResNet-50 | 11.0 | 49.3 | 10.4 | 7.3 | 16.0 | **3.7** | 6.5 |
| Adaptive-NMS [26] | ×1.3 | VGG-16 | 10.8 | 54.0 | 11.4 | 6.2 | - | - | - |
| BGCNet (ours) | ×1 | ResNet-50 | **9.4** | **45.9** | **9.0** | 6.4 | **14.0** | **3.7** | **5.5** |
| BGCNet (ours) | ×1 | HRNet-w32 | **8.8** | **43.9** | **8.0** | **6.1** | **11.6** | **2.6** | **5.3** |




(a) Reasonable (b) Heavy occlusion

**Figure 4: Comparisons of BGCNet with state-of-the-art methods on Caltech subsets: reasonable, and heavy occlusion.**

position of kernel elements. Some methods [3, 16] iteratively use RoI pooling or location refinement modules to improve the quality of true positives. We also tried to stack more BGC layers, but its performance is slightly decreased. This is reasonable because BGC captures the entirety of pedestrians, and one BGC is exactly right for the adaptation. Stacking BGC layers may mislead the network to capture too much information from backgrounds and nearby crowded pedestrians. Besides, only the center points of boxes are positive samples in BGCNet, which doesn't require the progressive method in anchor-based methods to remove the close false positives.

## 4.4 Comparison with State-of-the-art Detectors

**CityPersons.** We compare the results of BGCNet with the state of the arts on the validation set of CitypPersons. Table 5 shows that BGCNet achieves the new state-of-the-art results, with the significant improvement of 2% $MR^{-2}$, 5.4% $MR^{-2}$ and 5.4% $MR^{-2}$ on the reasonable, heavy occluded, and small scale subsets, respectively, compared to the most competitive results. Without any additional information, such as segmentation labels or visible bounding boxes, BGCNet still achieves the best results on all subsets with different occlusion and scale levels, except the bare occlusion subset, on which the results of the BGCNet is 0.2% $MR^{-2}$ lower than the OR-CNN [51] method which is tested with 1.3× scaled images. These results demonstrate the effectiveness of the proposed BGCNet.

**Caltech.** We also evaluate the proposed BGCNet method on the Caltech dataset and compare the results with the state-of-the-art pedestrian detectors. The reasonable and heavy subsets are used to evaluate the performance of the detectors under different cases. As shown in Figure 4, the proposed BGCNet achieves the new state of the art with 4.1% $MR^{-2}$ on the reasonable subsets, compared to the 4.5% of CSP, the 5.0% of RepLoss and 7.5% of SA-FasterRCNN. Moreover, BGCNet also achieves the best performance with 42.0% $MR^{-2}$ on the heavy occluded settings, which significantly outperforms the closest competitor CSP by 3.8% $MR^{-2}$.

Although BGCNet is designed for the upright pedestrian detection in scenarios of intelligent vehicles and surveillance, we further evaluate our method on the human detection dataset of CrowdHuman [37] which is collected from the Internet and contains many upper-body portraits. The detailed results on CrowdHuman are in the auxiliary material.

## 5 CONCLUSION

In this paper, we propose an effective and efficient pedestrian detector, BGCNet, to simultaneously address the fundamental challenges of various scales, occlusions and numerous false positives in pedestrian detection. The proposed BGC layer adaptively adjusts the receptive fields to capture the full information of various scale pedestrians, and places the kernel parameters around the head and foot to focus on the semantic and mostly visible features. A novel LM loss is introduced to constrain the relative magnitude of the neighbouring points in the score map to decrease the number of false positives. The extensive experiments on the CityPersons and Caltech datasets demonstrate the effectiveness of BGCNet. We expect other anchor-free detectors could also benefit from the proposed BGC layer and LM loss.

## ACKNOWLEDGMENTS

This work was partly supported by the NSFC Project #61672521. The authors would like to thank Anna Hennig who helped proofreading the paper.

## REFERENCES


[1] Zhaowei Cai, Quanfu Fan, Rogério Schmidt Feris, and Nuno Vasconcelos. 2016. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*. 354–370. https://doi.org/10.1007/978-3-319-46493-0_22

[2] Zhaowei Cai, Mohammad J. Saberian, and Nuno Vasconcelos. 2015. Learning Complexity-Aware Cascades for Deep Pedestrian Detection. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. 3361–3369. https://doi.org/10.1109/ICCV.2015.384

[3] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 6154–6162. https://doi.org/10.1109/CVPR.2018.00644

[4] Jiale Cao, Yanwei Pang, Jungong Han, and Xuelong Li. 2019. Hierarchical Shot Detector. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 9704–9713. https://doi.org/10.1109/ICCV.2019.00980

[5] Jiale Cao, Yanwei Pang, and Xuelong Li. 2019. Triply Supervised Decoder Networks for Joint Detection and Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 7392–7401. https://doi.org/10.1109/CVPR.2019.00757

[6] Yunyun Cao, Sugiri Pranata, Makoto Yasugi, Zhiheng Niu, and Hirofumi Nishimura. 2012. Stagged multi-scale LBP for pedestrian detection. In *19th IEEE International Conference on Image Processing, ICIP 2012, Lake Buena Vista, Orlando, FL, USA, September 30 - October 3, 2012*. 449–452. https://doi.org/10.1109/ICIP.2012.6466893

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2015. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.7062

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 3213–3223. https://doi.org/10.1109/CVPR.2016.350

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable Convolutional Networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 764–773. https://doi.org/10.1109/ICCV.2017.89

[10] Navneet Dalal and Bill Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*. 886–893. https://doi.org/10.1109/CVPR.2005.177

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. 248–255. https://doi.org/10.1109/CVPRW.2009.5206848

[12] Piotr Dollár, Ron Appel, Serge J. Belongie, and Pietro Perona. 2014. Fast Feature Pyramids for Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 8 (2014), 1532–1545. https://doi.org/10.1109/TPAMI.2014.2300479

[13] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. 2009. Pedestrian detection: A benchmark. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. 304–311. https://doi.org/10.1109/CVPRW.2009.5206631

[14] Zhenyu Duan, Jinpeng Lan, Yi Xu, Bingbing Ni, Lixue Zhuang, and Xiaokang Yang. 2017. Pedestrian Detection via Bi-directional Multi-scale Analysis. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*. 1023–1031. https://doi.org/10.1145/3123266.3123356

[15] Philip Geismann and Alois Knoll. 2010. Speeding Up HOG and LBP Features for Pedestrian Detection by Multiresolution Techniques. In *Advances in Visual Computing - 6th International Symposium, ISVC 2010, Las Vegas, NV, USA, November 29-December 1, 2010. Proceedings, Part I*. 243–252. https://doi.org/10.1007/978-3-642-17289-2_24

[16] Spyridon Gidaris and Nikos Komodakis. 2016. Attend Refine Repeat: Active Box Proposal Generation via In-Out Localization. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith (Eds.). BMVA Press. http://www.bmva.org/bmvc/2016/papers/paper090/index.html

[17] Ross B. Girshick. 2015. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. 1440–1448. https://doi.org/10.1109/ICCV.2015.169

[18] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *CoRR* abs/1706.02677 (2017). arXiv:1706.02677 http://arxiv.org/abs/1706.02677

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. 1026–1034. https://doi.org/10.1109/ICCV.2015.123

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[21] Li WangDong-Chen He. 1990. Texture classification using texture spectrum. *Pattern Recognition* 23, 8 (1990), 905–910. https://doi.org/10.1016/0031-3203(90)90135-8

[22] Van-Dung Hoang, My Ha Le, and Kang-Hyun Jo. 2014. Hybrid cascade boosting machine using variant scale blocks based HOG features for pedestrian detection. *Neurocomputing* 135 (2014), 357–366. https://doi.org/10.1016/j.neucom.2013.12.017

[23] Jan Hendrik Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. 2015. Taking a deeper look at pedestrians. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 4073–4082. https://doi.org/10.1109/CVPR.2015.7299034

[24] Hei Law and Jia Deng. 2018. CornerNet: Detecting Objects as Paired Keypoints. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*. 765–781. https://doi.org/10.1007/978-3-030-01264-9_45

[25] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 2999–3007. https://doi.org/10.1109/ICCV.2017.324

[26] Songtao Liu, Di Huang, and Yunhong Wang. 2019. Adaptive NMS: Refining Pedestrian Detection in a Crowd. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 6459–6468. http://openaccess.thecvf.com/content_CVPR_2019/html/Liu_Adaptive_NMS_Refining_Pedestrian_Detection_in_a_Crowd_CVPR_2019_paper.html

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

[28] Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. 2018. Learning Efficient Single-Stage Pedestrian Detectors by Asymptotic Localization Fitting. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*. 643–659. https://doi.org/10.1007/978-3-030-01264-9_38

[29] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. 2019. High-Level Semantic Feature Detection: A New Perspective for Pedestrian Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 5187–5196. http://openaccess.thecvf.com/content_CVPR_2019/html/Liu_High-Level_Semantic_Feature_Detection_A_New_Perspective_for_Pedestrian_Detection_CVPR_2019_paper.html

[30] Markus Mathias, Rodrigo Benenson, Radu Timofte, and Luc Van Gool. 2013. Handling Occlusions with Franken-Classifiers. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. 1505–1512. https://doi.org/10.1109/ICCV.2013.190

[31] Woonhyun Nam, Piotr Dollár, and Joon Hee Han. 2014. Local Decorrelation For Improved Pedestrian Detection. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 424–432. http://papers.nips.cc/paper/5419-local-decorrelation-for-improved-pedestrian-detection

[32] Wanli Ouyang and Xiaogang Wang. 2013. Joint Deep Learning for Pedestrian Detection. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. 2056–2063. https://doi.org/10.1109/ICCV.2013.257

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[34] Rabia Rauf, Ahmad R. Shahid, Sheikh Ziauddin, and Asad Ali Safi. 2016. Pedestrian detection using HOG, LUV and optical flow as features with AdaBoost as classifier. In *Sixth International Conference on Image Processing Theory, Tools and Applications, IPTA 2016, Oulu, Finland, December 12-15, 2016*. 1–4. https://doi.org/10.1109/IPTA.2016.7821024

[35] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 779–788. https://doi.org/10.1109/CVPR.2016.91

[36] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

[37] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. 2018. CrowdHuman: A Benchmark for Detecting Human in a Crowd. *CoRR* abs/1805.00123 (2018). arXiv:1805.00123 http://arxiv.org/abs/1805.00123

[38] Tao Song, Leiyu Sun, Di Xie, Haiming Sun, and Shiliang Pu. 2018. Small-scale Pedestrian Detection Based on Somatic Topology Localization and Temporal Feature Aggregation. *CoRR* abs/1807.01438 (2018). arXiv:1807.01438 http://arxiv.org/abs/1807.01438

[39] Tao Song, Leiyu Sun, Di Xie, Haiming Sun, and Shiliang Pu. 2018. Small-Scale Pedestrian Detection Based on Topological Line Localization and Temporal Feature Aggregation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*. 554–569. https://doi.org/10.1007/978-3-030-01234-2_33

[40] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 5693–5703. http://openaccess.thecvf.com/content_CVPR_2019/html/Sun_Deep_High-Resolution_Representation_Learning_for_Human_Pose_Estimation_CVPR_2019_paper.html

[41] Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. https://openreview.net/forum?id=ry8u21rtl

[42] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Strong Parts for Pedestrian Detection. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. 1904–1912. https://doi.org/10.1109/ICCV.2015.221

[43] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *Proc. Int. Conf. Computer Vision (ICCV)*.

[44] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. 2013. Selective Search for Object Recognition. *International Journal of Computer Vision* 104, 2 (2013), 154–171. https://doi.org/10.1007/s11263-013-0620-5

[45] Paul A. Viola and Michael J. Jones. 2001. Rapid Object Detection using a Boosted Cascade of Simple Features. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), with CD-ROM, 8-14 December 2001, Kauai, HI, USA*. 511–518. https://doi.org/10.1109/CVPR.2001.990517

[46] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. 2018. Repulsion Loss: Detecting Pedestrians in a Crowd. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 7774–7783. https://doi.org/10.1109/CVPR.2018.00811

[47] Mo Zhang, Jie Zhao, Xiang Li, Li Zhang, and Quanzheng Li. 2019. ASCNet: Adaptive-Scale Convolutional Neural Networks for Multi-Scale Feature Learning. *CoRR* abs/1907.03241 (2019). arXiv:1907.03241 http://arxiv.org/abs/1907.03241

[48] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. 2017. Scale-Adaptive Convolutions for Scene Parsing. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2050–2058. https://doi.org/10.1109/ICCV.2017.224

[49] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hendrik Hosang, and Bernt Schiele. 2016. How Far are We from Solving Pedestrian Detection?. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 1259–1267. https://doi.org/10.1109/CVPR.2016.141

[50] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. 2017. CityPersons: A Diverse Dataset for Pedestrian Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 4457–4465. https://doi.org/10.1109/CVPR.2017.474

[51] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. 2018. Occlusion-Aware R-CNN: Detecting Pedestrians in a Crowd. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*. 657–674. https://doi.org/10.1007/978-3-030-01219-9_39

[52] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. 2018. Single-Shot Refinement Neural Network for Object Detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 4203–4212. https://doi.org/10.1109/CVPR.2018.00442

[53] Chunluan Zhou and Junsong Yuan. 2016. Learning to Integrate Occlusion-Specific Detectors for Heavily Occluded Pedestrian Detection. In *Computer Vision - ACCV 2016 - 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*. 305–320. https://doi.org/10.1007/978-3-319-54184-6_19

[54] Chunluan Zhou and Junsong Yuan. 2018. Bi-box Regression for Pedestrian Detection and Occlusion Estimation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*. 138–154. https://doi.org/10.1007/978-3-030-01246-5_9

[55] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krähenbühl. 2019. Bottom-Up Object Detection by Grouping Extreme and Center Points. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 850–859. http://openaccess.thecvf.com/content_CVPR_2019/html/Zhou_Bottom-Up_Object_Detection_by_Grouping_Extreme_and_Center_Points_CVPR_2019_paper.html

[56] Chenchen Zhu, Yihui He, and Marios Savvides. 2019. Feature Selective Anchor-Free Module for Single-Shot Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 840–849. http://openaccess.thecvf.com/content_CVPR_2019/html/Zhu_Feature_Selective_Anchor-Free_Module_for_Single-Shot_Object_Detection_CVPR_2019_paper.html