Self-Mimic Learning for Small-scale Pedestrian Detection

Jialian Wu[1], Chunluan Zhou[2], Qian Zhang[3], Ming Yang[3] and Junsong Yuan[1]
[1]State University of New York at Buffalo
[2]Wormpex AI Research [3]Horizon Robotics, Inc.
{jialianw,jsyuan}@buffalo.edu

ABSTRACT

Detecting small-scale pedestrians is one of the most challenging problems in pedestrian detection. Due to the lack of visual details, the representations of small-scale pedestrians tend to be weak to be distinguished from background clutters. In this paper, we conduct an in-depth analysis of the small-scale pedestrian detection problem, which reveals that weak representations of small-scale pedestrians are the main cause for a classifier to miss them. To address this issue, we propose a novel Self-Mimic Learning (SML) method to improve the detection performance on small-scale pedestrians. We enhance the representations of small-scale pedestrians by mimicking the rich representations from large-scale pedestrians. Specifically, we design a mimic loss to force the feature representations of small-scale pedestrians to approach those of large-scale pedestrians. The proposed SML is a general component that can be readily incorporated into both one-stage and two-stage detectors, with no additional network layers and incurring no extra computational cost during inference. Extensive experiments on both the CityPersons and Caltech datasets show that the detector trained with the mimic loss is significantly effective for small-scale pedestrian detection and achieves state-of-the-art results on CityPersons and Caltech, respectively.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**.

KEYWORDS

pedestrian detection; small-scale object detection; mimic learning

ACM Reference Format:
Jialian Wu, Chunluan Zhou, Qian Zhang, Ming Yang and Junsong Yuan. 2020. Self-Mimic Learning for Small-scale 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, 9 pages. https://doi.org/10.1145/3394171.3413634

1 INTRODUCTION

Pedestrian detection is the cornerstone for many vision applications, such as autonomous driving, robotics and video surveillance. In recent years, advances in convolutional neural networks (CNNs) and object detection have significantly boosted the performance for pedestrian detection [2, 13, 18, 20–24, 31, 32, 39, 40, 43, 44]. Current detectors perform well on large-scale pedestrians, yet limited accuracies on small-scale ones due to low-resolution figures and downsampling operations. In fact, small-scale pedestrians occur quite frequently in many scenarios of pedestrian detection. For example, in the Caltech [7] dataset, 69% of pedestrians have heights of 30-80 pixels, and 15% of pedestrians are under 30 pixels. It is indeed critical to detect such small-scale pedestrians at a distance for various application scenarios. For example, a pedestrian far away in front of a self-driving car should be detected early enough for the control system to avoid any chance of collision reliably and smoothly.

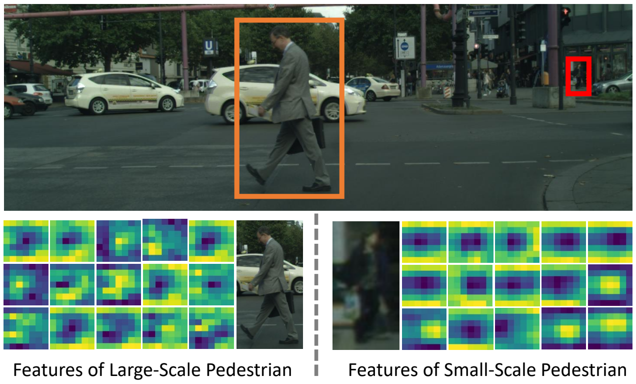

Features of Large-Scale Pedestrian
Features of Small-Scale Pedestrian

Figure 1: Visualization of RoI feature maps from a large-scale and a small-scale pedestrians. In our implementation, the RoI features of a pedestrian have $7 \times 7 \times 256$ dimensions. Each feature map shown above is selected from the 256 feature channels.

Recently, some efforts [14, 16] attempt to address the small-scale object detection problems via exploiting rich representations of large-scale objects. Li *et al.* [16] propose a perceptual generative adversarial network (Perceptual GAN) to generate super-resolved representations for small-scale objects. While this approach enhances the representations of small-scale objects, the generator used in the perceptual GAN introduces relatively high computational overhead for producing super-resolved representations during inference. To reduce feature variations among objects of different scales, Kim *et al.* [14] propose a Scale Aware Network (SAN) to map convolutional features from different scales to a scale-invariant subspace. However, the SAN simply upsamples image patches of small objects, which could result in blurred image patches. In this paper, we aim to design a simple and effective approach to enhance feature representations for small-scale pedestrians with minimal computation overhead in inference.

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

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413634

Inspired by mimicking techniques used in model acceleration and compression [1, 12, 29], we exploit a mimic method to enhance the representations of small-scale pedestrians. The essence of mimic learning in object detection [17, 33] is that a small model can learn better representations by mimicking the features from a large model. In view of this, we propose a self-mimic learning method, which extends the mimicking techniques to learn super-resolved representations for small-scale pedestrians with the help of large-scale pedestrians in a *single* model, therefore achieving self-mimic learning in the model. Specifically, we train a deep CNN via a mimic loss, which aims at forcing the feature distribution of small-scale pedestrians to approximately mimic that of large-scale pedestrians from the same network architecture.

We implement our mimic method based on the instance-level features. As shown in Figure 1, the RoI features (output by the RoI Align [10]) of large-scale pedestrians generally retain more fertail information than those of small-scale pedestrians. We therefore use them as the reference to help small-scale pedestrians approximately learn the feature distribution of large-scale pedestrians. By forcing the features of small-scale pedestrians to approach those of large-scale pedestrians, it leads to two benefits for small-scale pedestrian detection. First, the missing details of small-scale pedestrians are compensated in feature space, therefore enhancing the representations of small-scale pedestrians. Second, the intra-class variances of pedestrian features are reduced, which makes it easier for a classifier to distinguish small-scale pedestrians from backgrounds.

The proposed SML is a general component that can be readily incorporated into both one-stage and two-stage detectors with any backbone network. It only imposes supervision from large-scale pedestrians on small-scale pedestrians via a mimic loss without adding any network layers. Therefore, it does not incur any additional computational cost during inference. To validate the effectiveness of SML, we conduct thorough experiments on the Caltech [7] and CityPersons [38] pedestrian detection datasets using ResNet-18 and ResNet-50 [11], respectively. SML effectively improves the performance on small-scale pedestrians and achieves state-of-the-art detection performance.

2 RELATED WORK

Pedestrian Detection. With the renaissance of deep learning, convolutional neural networks (CNN) based pedestrian detectors have achieved promising performance in recent years [3–5, 24–26, 35, 36, 41, 42, 45, 46]. However, occluded pedestrian detection and small-scale pedestrian detection remain two challenges for most existing methods. To handle occlusions, recent works are designed by exploiting attention mechanism [27, 40] and part-based detection [24, 39, 43]. To alleviate small-scale detection problems, Song *et al.* [31] utilize somatic topology lines to locate pedestrians, which is helpful for localizing small-scale pedestrians. In this paper, we propose a novel self-mimic learning method for enhancing the feature representations of small-scale pedestrians.

Small-scale Object Detection. Although current detectors achieve exceptional performance on large-scale objects [8, 9, 28, 34], they can not handle small-scale objects very well. In the past decade, many efforts attempt to tackle the problem of large scale variations across object instances at different scales [14, 19, 30, 47]. Lin *et al.* [19] propose a top-down architecture to combine shallow layers with high-level semantic feature maps for multi-scale object detection. Li *et al.* [16] propose a perceptual generative adversarial network (Perceptual GAN) to transfer the representation of small-scale objects to super-resolved ones. Nevertheless, the perceptual GAN needs a heavy generator to generate a residual representation which is then added to the original small-scale object features such that the small-scale objects can be classified by a discriminator as large-scale objects. Therefore, a relatively large computational cost is incurred during inference. Kim *et al.* [14] propose a Scale Aware Network (SAN) to learn the relationship among the objects at different scales. The SAN may result in blurred image patches when it upsamples image regions of small objects, and does not explore the relationship among different instances within the same class. In contrast, our proposed SML exploits high resolution pedestrian examples to enhance the representations of small-scale pedestrians with no extra computation in inference. Also, SML effectively minimizes the intra-class variances among pedestrians at various scales.

Mimicking Techniques. Network mimicking or distilling is recently used for model acceleration and compression. The assumption of network mimicking is that a small model can learn better representations by adding supervision from a large model. As pioneered in [1, 12], knowledge distillation are used for simple classification tasks by requiring a small model to mimic the logits of a large model. Li *et al.* [17] enhance the representations of a small model by mimicking the features sampled from a large model, which first extends mimic methods to object detection tasks. Wei *et al.* [33] propose a quantization mimic in order to shrink the search scope for small model. Unlike these methods, we apply mimicking techniques in a *single model* by making the features of small-scale pedestrians mimic those of large-scale pedestrians from the same network architecture. Recently, Li *et al.* [15] also exploit the idea of self-mimicking to improve the small-scale object detection. In [15], a feature intertwiner is introduced into the two-stage detector, Faster R-CNN, to improve small-scale object detection. The feature intertwiner transforms the features of small-scale objects to make them approach the global centroid of the features of large-scale objects. In contrast, our method utilizes local centroids of large-scale pedestrian features to guide the mimic learning, which is more effective for the pedestrian detection task (as evidenced in Table 2). Moreover, we also discuss the localization problem caused by mimic learning and extend the proposed SML to one-stage detectors.

3 ANALYSIS OF SMALL-SCALE PEDESTRIAN DETECTION

Dataset. Small-scale pedestrians occur quite frequently in real-world scenes. We conduct an analysis on the CityPersons [38] dataset which contains $5,000$ images with $35,000$ persons and $13,000$ ignored region annotations. The Reasonable subset is a common setup for evaluating comprehensive performance of a pedestrian detector, in which pedestrians are taller than 50 pixels and not occluded more than 35%. Among them, small-scale pedestrians still account for a large percentage. For example, the percentage

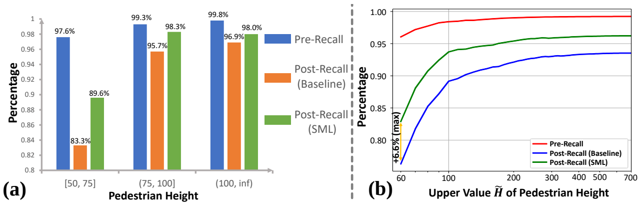


Figure 2: (a) Recall histogram of pedestrians at different scales. Pedestrians with heights in $[50, 75]$ pixels are defined as small-scale pedestrians. (b) Recall curves of pedestrians at different scales. Each point on the curves represents the recall calculated in the height interval of $[50, \widetilde{H})$ pixels.

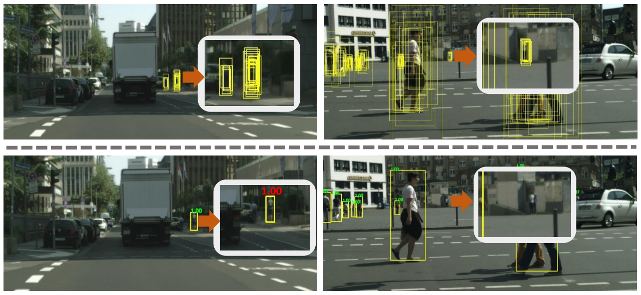


Figure 3: Examples of wrong classification in the baseline detector. (Top) Proposals that is generated by the RPN in two images. The proposals with low scores have been removed. (Bottom) Final detection results where boxes are eliminated if their classification scores are lower than 0.5.

of pedestrians with heights in $[50, 75]$ is 24.3%, while the resolutions of original images are $1,024 \times 2,048$ pixels.

Missed Detection. We take the two-stage detector as an example to analyze the misclassification problem of small-scale pedestrians. In the two-stage detector, *e.g.,* Faster R-CNN [28], there are two phases for detecting an object. The first phase is to generate proposals and use them to extract the RoI features. In the second phase, these RoI features are fed into the head network for classification and bounding box regression. To cover objects at different scales, the region proposal network (RPN) generates a large number of proposals in the first phase. Nevertheless, the detector still tends to miss small-scale pedestrians in the second phase even with massive proposals.

To delve into the missed detection problem of small-scale pedestrians, we calculate two types of recalls, *Pre-Recall* and *Post-Recall*, for pedestrians at different scales. We define the *Pre-Recall* as the percentage of ground-truth examples which are associated with at least one proposal and the *Post-Recall* as the percentage of ground-truth examples which are associated with at least one correctly classified proposal (i.e., classification score ≥ 0.5). We use the Faster R-CNN based detector that is our baseline (described in Section 4.1) for this analysis, in which $1,000$ proposals are generated from the RPN in our experiment. As shown in Figure 2 (a), the *Pre-Recalls* for pedestrians of different scales are close and reach a high value of over 95%. However, the recall of the baseline notably drops from 97.6% (*Pre-Recall*) to 83.3% (*Post-Recall*) for small-scale pedestrians, while the declines from *Pre-Recall* to *Post-Recall* for large-scale pedestrians are only 3.6% and 2.9%, respectively. We can also see in Figure 2 (b) that there is a large gap between the *Pre-Recall* and *Post-Recall* when $\widetilde{H}$ is less than 100 pixels. The results in Figure 2 indicate that even though the majority of small-scale pedestrians are detected by the RPN, many of them are not correctly classified in the second phase, probably due to the weak feature representations.

To better understand the missed detection problem of small-scale pedestrians, some missed detection examples are visualized in Figure 3. We observe that the small-scale pedestrians are located with accurate proposals generated by the RPN. However, they are regarded as background regions with classification scores lower than 0.5 by the head network. The analysis in this section indicates that the misclassification mainly causes the poor performance on small-scale pedestrians.

4 SELF-MIMIC LEARNING

In this section, we first provide an overview of our method based on the two-stage detector, and then present the mimic loss function. Afterwards, we introduce the implementation schemes of SML. We also extend our method to one-stage detector as presented in supplementary material.

4.1 Overview

Baseline Detector. In this paper, we use Faster R-CNN [28] as the baseline detector and implement it using ResNet-18 or ResNet-50 backbones with feature pyramid network (FPN) [19] and deformable convolution [6].

Self-Mimic Learning. In previous works [1, 12, 17, 29, 33], mimicking techniques are mostly used in model acceleration and compression. We extend the mimic idea to feature space of data samples from the same neural network, enabling the feature distribution of small-scale pedestrians to approach that of the larger ones.

According to the analysis in Section 3, most regions of small-scale pedestrians can be well located by the proposals generated from the RPN. Considering large-scale pedestrians usually have richer feature representations, it is a reasonable idea to enforce the distribution of small-scale pedestrians to mimic that of large-scale pedestrians in the RoI feature space, such that the representations of small-scale pedestrians are enhanced and conceptually encoded with more visual details as pedestrians with higher resolutions. In SML, pedestrians with heights in $(0, H_S]$ pixels are defined as small-scale pedestrians, and those with heights in $(H_S, H_L]$ pixels are regarded as large-scale pedestrians. We do not consider the super large-scale pedestrians whose heights are taller than H_L pixels since there is a large gap between the representations of super large-scale and small-scale pedestrians.

Figure 4 (a) illustrates the framework of our approach. An input image is first fed into the feature pyramid network (FPN) to extract multi-scale feature maps. Then, the RPN is built upon these feature maps to generate proposals with various sizes. The RoI Align operation takes the proposals and feature maps as input and outputs RoI features with dimensions of $7 \times 7 \times 256$. Among these RoI features, the features of small-scale pedestrians mimic those of large-scale

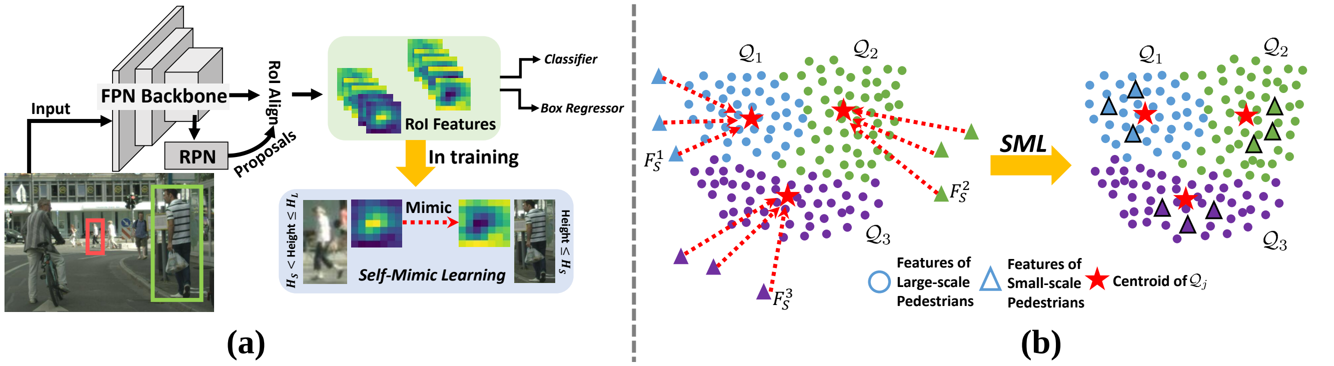


Figure 4: (a) The proposed architecture of SML. An input image is fed into the FPN backbone, and then the proposals generated by the RPN are used for extracting RoI features by RoI Align. At training step, the RoI features of small-scale pedestrians are trained to mimic those of large-scale pedestrians. During inference, the detection pipeline of SML is the same as that of the baseline detector without any extra computational cost. (b) Illustration of our proposed mimic loss.

pedestrians. By optimizing the mimic loss (defined in Section 4.2), not only small-scale pedestrians learn super-resolved representations from the large-scale ones, but also the intra-class variance can be reduced. Subsequently, with the richer and more consistent representations, the RoI features of small-scale pedestrians are fed into the head network for easier classification and regression.

4.2 Mimic Loss

For detection tasks, feature maps are key to determine both classification and localization accuracy. We implement the proposed mimic method on the RoI features of the small-scale and large-scale pedestrians. The overall loss function for SML is defined as:

$$L = L_{cls}^{R} + L_{box}^{R} + L_{cls}^{H} + L_{box}^{H} + \alpha \times L_m, \tag{1}$$

where L_{cls}^{R} and L_{box}^{R} are the classification loss and bounding box regression loss respectively for the RPN, and L_{cls}^{H} and L_{box}^{H} are the classification loss and bounding box regression loss respectively for the head network. For the classification losses L_{cls}^{R} and L_{cls}^{H}, we use the cross-entropy loss over two classes (pedestrian and background). For the regression losses L_{box}^{R} and L_{box}^{H}, we use the loss function Smooth$_{L1}$ described in [8]. L_m denotes the mimic loss that is weighted by α.

We denote by $F_L \in \mathcal{R}^d$ the RoI features of large-scale pedestrians ($H_S < height \le H_L$) and $F_S \in \mathcal{R}^d$ the RoI features of small-scale pedestrians ($0 < height \le H_S$), respectively, and $d = 7 \times 7 \times 256$ is the feature dimensionality. Note that here a pedestrian refers to a proposal which has an IoU ≥ 0.5 with one ground-truth pedestrian example in the training image. In the training data, we denote $\mathcal{L} = \{F_L^1, ..., F_L^N\}$ the RoI features of N large-scale pedestrians and $\mathcal{S} = \{F_S^1, ..., F_S^M\}$ the RoI features of M small-scale pedestrians. Ideally, our goal is to force the features in $\mathcal{S}$ to mimic those in $\mathcal{L}$ such that $\mathcal{S}$ and $\mathcal{L}$ have similar probability distributions in feature space, *i.e.*, $p(F_S) \approx p(F_L)$. However, it is difficult to optimize the high-dimensional distribution directly given limited training examples. In general, the distribution of pedestrian features is multi-modality, we therefore choose to approximately achieve the mimic learning goal by pushing the features of each small-scale pedestrian in $\mathcal{S}$ to one of the local centroids of the features of large-scale pedestrians in $\mathcal{L}$. Thus, we define the mimic loss L_m as:

$$\begin{aligned} L_m &= \frac{1}{M}\sum_{j=1}^{M}\left\| F_S^j - \frac{1}{|Q_j|}\sum_{F_G \in Q_j} F_G \right\|_2 \\ &= \frac{1}{M}\sum_{j=1}^{M}\left\| F_S^j - C_j \right\|_2 \end{aligned} \tag{2}$$

For each small pedestrian feature F_S^j, $Q_j \subseteq \mathcal{L}$ is a set of its mimic features in $\mathcal{L}$. When $Q_j \subset \mathcal{L}$, Q_j can be regarded as a sub modality of all the large-scale pedestrian samples in the training set, and C_j is a local centroid in $\mathcal{L}$. Conversely, when $Q_j = \mathcal{L}$, C_j becomes the global mean feature of large-scale pedestrians in $\mathcal{L}$. Figure 4 (b) illustrates the idea of Eq. 2. We introduce how to select the local centroids C_j for each F_S^j in the next section.

4.3 Implementation

We introduce two implementation schemes to select the local centroid C_j for each small-scale pedestrian sample as follows:

Offline self-mimic learning. In this scheme, we first train a baseline detector on the dataset as the reference detector, and collect all the RoI features of large-scale pedestrians. Then, we use k-means to cluster these RoI features into K clusters. We compute the distances between the RoI features of each small-scale pedestrian F_S^j and the center of each cluster. We choose the nearest cluster center as C_j for F_S^j. The Offline SML is a straightforward way to partition the feature space of large-scale pedestrians and employ their centroids to guide the feature learning of small-scale pedestrians.

Online self-mimic learning. In this scheme, we only train the network for once without training an additional reference detector. For each F_S^j, we define the local centroid C_j as the average of the RoI features of large-scale pedestrians in the same image where F_S^j is obtained. During training, we do not back-propagate the gradients of the mimic loss for large-scale pedestrians, because the large-scale pedestrians only serve as references for small-scale

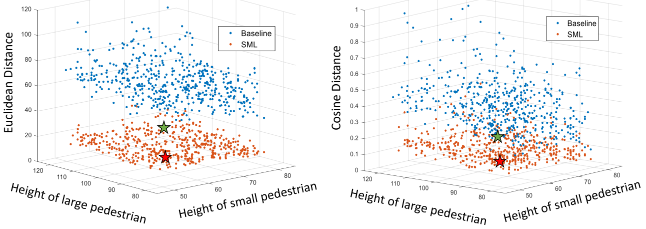


Figure 5: Feature distances of the RoI features between small-scale and large-scale pedestrians on the CityPersons validation set. Each point represents a pair of small-scale and large-scale pedestrians. The green and red stars show the distances between the mean feature of small-scale pedestrians and that of large-scale pedestrians of the baseline and SML, respectively.

pedestrians to learn the super-resolved representations. According to the statistics of pedestrian detection dataset, there are over 75% images that contain both small-scale and large scale pedestrian examples. Thus, the Online SML is efficient and easy to implement the mimic learning in an end-to-end fashion.

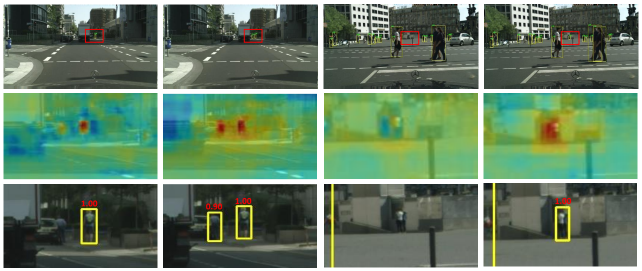
The Offline SML considers the different modalities of visual features of all the large-scale pedestrians samples. While the Online SML focus more on the modalities of context of images (*e.g.*, illumination and weather condition). We discuss and compare these two schemes in Section 6.2.

5 DISCUSSION

The SML brings two benefits: 1) enhancing representations of small pedestrians and 2) reducing intra-class variance, together contributing to the performance improvement on small-scale pedestrians as well as the overall performance improvement on all scales.

1) Representation enhancement for small-scale pedestrians: The mimic loss forces the features of small-scale pedestrians to approach the local feature centroids of large-scale pedestrians, so as to compensate their missing details to some extent in feature space. This effectively enhances the feature representations of small-scale pedestrians, therefore improving the detection performance on small-scale pedestrians. To verify this, we use the same set of proposals as the baseline detector and calculate the *Post-Recall* for SML. As shown in Figure 2 (1), SML outperforms the baseline detector with a remarkable improvement of 6.3% in *Post-Recall* on small-scale pedestrians, achieving a better classification for small-scale pedestrians. In Figure 2 (b), SML consistently improves the *Post-Recall* over the baseline by a large margin in different height ranges, and obtains a maximum gain of 6.6% in the height interval [50, 60]. Figure 6 shows qualitative comparison between the baseline and SML. We can see that with SML the regions of small-scale pedestrians in the feature maps are more discriminative and have higher responses due to the feature compensation of SML. These quantitative and qualitative analyses validate the effect of the SML for enhancing feature representations for small-scale pedestrians.

2) Intra-class variance reduction: We calculate the feature distances between small-scale and large-scale pedestrians. We use two types of distance metrics: Euclidean distance and cosine distance. The Euclidean distance, which is defined as $D_{Euclidean} = \|F_L - F_S\|_2$, describes the spatial distance of RoI features between two instances. The cosine distance, which is calculated as $D_{consine} = 1 - \frac{F_L \cdot F_S}{\|F_L\|\|F_S\|}$, describes the orientation similarity of RoI features between two instances. As shown in Figure 5, SML significantly reduces both the Euclidean distances and cosine distances between the small-scale and large-scale pedestrians in feature space. As a result, intra-class variance is reduced, making it easier for the head network to classify pedestrians and backgrounds. It also benefits the detection of large-scale pedestrians as shown in Figure 2.



(a) Baseline (b) SML (c) Baseline (d) SML

Figure 6: Qualitative comparison of the baseline and SML. The first row shows the detection results in the original images. The second and third rows show the feature maps and detection results respectively of the red boxes. The feature maps are selected from the P2 in FPN. Red boxes denote areas where the baseline detector can not correctly classify tiny pedestrians. We remove boxes whose classification scores are less than 0.5. More visualization results can be found in supplementary material.

Table 1: Subset settings in the Caltech and CityPersons datasets.

Dataset	Training	Evaluation
Caltech (1× training set)	$h \in [25, \infty]$ $v \in [0.65, 1]$	Reasonable, Far, Medium
Caltech (10× training set)	Reasonable	Reasonable, Small
CityPersons	Reasonable	Reasonable, Small, Partial, Bare

6 EXPERIMENTS

6.1 Datasets and Experiments Settings

Caltech Dataset. The Caltech [7] dataset consists of 11 sets of videos. The training set is divided into two subsets: 1× and 10× training sets, which sample 4, 250 and 42, 782 frames from the first six sets, respectively. For evaluation, the testing set samples 4, 024 frames from the last five sets. The original images are upsampled to $960 \times 1,280$ pixels. In our experiments, the new annotations with high quality provided by [37] are used for both training and testing. We find most pedestrians whose heights are less than 50 pixels are ignored in the new annotations of the 10× training sets. Therefore,

Table 2: Comparisons of the Online SML and the Offline SML.

Dataset	Subset	Online	Offline SML			
		SML	1(K)	20(K)	40(K)	60(K)
Caltech	Reasonable	6.8	7.7	7.7	6.7	7.2
	Medium	21.2	23.2	22.8	22.3	23.1
	Far	63.6	64.9	63.4	63.5	62.5
CityPerson	Reasonable	12.3	13.3	12.9	12.1	12.3
	Small	19.3	21.1	19.4	18.0	20.3

we use the 1× training set for demonstrating the effectiveness of SML on small-scale pedestrians and the 10× training sets for comparison with the state-of-the-art methods. For 1× training set, we train the network for $80k$ iterations. The initial learning rates are set to 0.0075 and 0.001 for ResNet-18 and ResNet-50, respectively, and decreased by a factor of 10 after the first $60k$ iterations. For 10× training set, the learning rate is set to 0.001 for the first $200k$ iterations and 0.0001 for the next $100k$ iterations.

CityPersons Dataset. The CityPersons [38] dataset is introduced in Section 3. We train our SML detector on the CityPersons training set (2975 images) and evaluate the performance on the validation set (500 images). The learning rate is set to 0.004 for the first $60k$ iterations and 0.0004 for the rest of $30k$ iterations. The scales of original images are $1,024 \times 2,048$ pixels.

Experiments Settings. In Table 1, we list different subset settings for training and evaluation, which follows the common setup used in the Caltech and CityPersons datasets. These subsets are defined as: Reasonable: $h \in [50, \infty]$, $v \in [0.65, 1]$; Far: $h \in [0, 30]$, $v \in [0.65, 1]$; Medium: $h \in (30, 80]$, $v \in [0.65, 1]$; Small: $h \in (50, 75]$, $v \in [0.65, 1]$; Partial: $h \in [50, \infty]$, $v \in [0.65, 0.9]$; Bare: $h \in [50, \infty]$, $v \in [0.9, 1]$, where h and v denote the height and visible ratio of pedestrian examples, respectively. Our baseline detector is described in Section 4.1. We adopt the standard evaluation metric in pedestrian detection: MR^{-2} (lower is better). According to the definition of the Medium subset in the Caltech dataset, we respectively set H_S and H_L to 80 and 120 by default. Also, we use the same settings of H_S and H_L for the CityPersons dataset. The mimic loss weight α is empirically set to 16 by default. We optimize SML using the Stochastic Gradient Descent (SGD) solver, which is trained on 4 Titan X GPUs with the mini-batch involving 1 image per GPU.

6.2 Ablation Study

In this section, we analyze different design choices of SML on both the Caltech and CityPersons datasets, and the 1× training set is used for the analyses on the Caltech dataset. We use ResNet-50 as the backbone network. The baseline detector achieves 7.7%/23.1%/65.2% MR^{-2} on the Reasonable/Medium/Far subsets of the Caltech dataset and 13.4%/21.9% MR^{-2} on the Reasonable/Small subsets of the CityPersons dataset, respectively.

Offline SML vs. Online SML. We compare the Online SML and the Offline SML with different numbers of clusters (*i.e.,* K) in Table 2. For the Offline SML, when $K = 1$, the mean of the RoI features of large-scale pedestrians (*i.e.,* $Q_j = \mathcal{L}$) is used to guide the feature learning of small-scale pedestrians. We see from the results that the Offline SML with $K = 1$ gets outperformed by both the Offline SML with $K > 1$ and the Online SML. We argue that the reason is that the distribution of pedestrian features is multi-modality. In other words, the global centroid of large-scale pedestrians in feature space may not well represent large-scale pedestrian examples. Thus, it validates the effectiveness of our approach for enforcing the small-scale pedestrian features to mimic the local centroids of the large-scale pedestrian features. Moreover, the results suggest that the Online SML achieves similar performance to the Offline SML, showing the effectiveness of both proposed schemes. Since the Online SML is more efficient and simple to be trained, we use the Online SML for analyzing our method in the remaining sections of the paper.

Table 3: Results of SML with different mimic schemes.

Dataset	Mimic Scheme	Feature-wise	Channel-wise
Caltech	Reasonable	**6.8**	7.6
	Medium	**21.2**	22.7
	Far	**63.6**	**63.6**
CityPersons	Reasonable	**12.3**	13.6
	Small	**19.3**	20.6

Table 4: Results of SML with different parameters β. When $\beta = \infty$, it means that we do not restrict the sampling scheme.

β IoU	Dataset	Subset	0.2	0.3	0.4	0.5	∞
0.5	Caltech	Reasonable	7.3	7.3	7.2	7.6	**6.8**
		Medium	22.2	22.7	22.2	22.5	**21.2**
		Far	64.4	64.7	64.9	64	**63.6**
	CityPersons	Reasonable	12.7	13.1	12.8	12.8	**12.3**
		Small	20.1	20.3	20.8	19.4	**19.3**
0.75	Caltech	Reasonable	27.7	28.05	**27.4**	28.4	28.2
		Medium	50.3	50	**48.7**	49.6	50
		Far	88.1	87.6	87.7	87.1	**86.3**
	CityPersons	Reasonable	36.8	36.2	**35.0**	35.4	36.2
		Small	47.8	50.4	47.6	**47.1**	48.9

Mimic Scheme. In SML, we minimize the Euclidean distances between the RoI features of small-scale and large-scale pedestrians (See Eq. 2). Similar to SAN [14], we also implement its channel-wise mimic scheme in our approach. Specifically, we apply global average pooling to each channel of the RoI features to obtain channel-wise features, and then minimize the distances between the channel-wise features of small-scale and large-scale pedestrians. As shown in Table 3, the channel-wise mimic scheme gets outperformed by the feature-wise mimic scheme, which indicates our feature-wise mimic scheme learns representations of higher quality than the channel-wise mimic scheme.

Sampling Scheme for Online SML. In the Online SML, Q_j is sampled from all the RoI features of large-scale pedestrians in the same image where F_S^j is obtained. The relative locations of small-scale and large-scale pedestrian proposals to their ground-truth boxes may not be consistent, which could result in poor offset predictions by the bounding box regression branch. Therefore, a

Table 5: Results of SML with different mimic loss functions.

Dataset	Mimic Loss Function	Euclidean	Cosine
Caltech	Reasonable	6.8	**6.7**
	Medium	**21.2**	22.3
	Far	63.6	**62.6**
CityPersons	Reasonable	**12.3**	**12.3**
	Small	19.3	**18.7**

Table 6: Comparison with the scale-based focal loss on the CityPersons dataset.

Method	SML	Scale-based Focal Loss	Baseline
Reasonable	**12.3**	13.6	13.4
Small	**19.3**	21.3	21.9

Table 7: Comparisons with the baseline detector on the Caltech testing set. 1× training set is used for training.

Backbone	Method	Reasonable	Medium	Far
ResNet-18	Baseline	7.9	24.5	65.3
	SML	**7.4**	**23.2**	**64.1**
ResNet-50	Baseline	7.7	23.1	65.2
	SML	**6.8**	**21.2**	**63.6**

more probably reasonable sampling scheme is to choose large-scale pedestrians which have similar relative locations to match small-scale pedestrians. To study the impact of sampling schemes on classification and localization, we design and add a restriction to the original sampling scheme, which is defined as:

$$\sqrt{(t_x^S - t_x^L)^2 + (t_y^S - t_y^L)^2} \quad \leq \beta, \tag{3}$$

where β is a restrictive parameter to ensure the relative locations of small-scale and large-scale pedestrians is similar. t^S and t^L are the normalized coordinate offsets relative to the ground-truth boxes of small-scale proposals and large-scale proposals, respectively. t_x and t_y which denote the offsets of x and y coordinates are calculated as:

$$t_x = \frac{x_{proposal} - x_{gt}}{w_{proposal}}, \quad t_y = \frac{y_{proposal} - y_{gt}}{h_{proposal}}. \tag{4}$$

In the new scheme, Q_j consists of large-scale pedestrian examples which satisfy Eq. 3. As shown in Table 4, under the IoU criterion of 0.5, the SML with the original sampling scheme ($\beta = \infty$) clearly improves detection performance over the baseline especially on the small-scale pedestrians, while it achieves less improvement under the IoU criterion of 0.75. With the new sampling scheme, the SML can achieve better performance under the IoU criterion of 0.75. It shows that SML with the restrictive sampling scheme can reduce negative effect of the original sampling scheme on the localization ability. Therefore, the new sampling scheme can serve as an alternative for SML when higher localization accuracy is required.

Mimic Loss Function. As shown in Table 5, we also experiment with the cosine distance instead of the Euclidean distance as the mimic loss function. The experimental results on the Caltech and CityPersons datasets indicate that both the cosine distance and Euclidean distance are effective loss functions for self-mimic learning.

SML vs. Scale-based Focal Loss. To improve the detection performance on the small-scale pedestrian examples, another possible solution is to increase their loss weights during training (*i.e.,* scale-based focal loss). Therefore, we further compare the scale-based focal loss with our proposed SML. Specifically, for the scale-based focal loss, we define the classification loss for pedestrian class as follows:

$$L_{cls}^H = \begin{cases} -(1 + \frac{H_S - h}{H_S - H_{min}}) \times \log p, & h \leq H_S \\ -\log p, & \text{otherwise} \end{cases} \tag{5}$$

where $p \in [0, 1]$ is the possibility predicted by the head network for the pedestrian class. h is the height of current pedestrian proposal and H_{min} is the minimum height of all the pedestrian proposals in the current mini-batch. As shown in Table 6, the proposed SML outperforms the baseline with scale-based focal loss, and the scale-based focal loss fails to improve performance on the Reasonable subset that contains large-scale pedestrians as well. We argue that the performance gap may caused by two reasons. First, it may be hard to compensate the missing details of small-scale pedestrians in feature space by simply increasing their loss weights. By contrast, our method exploit richer feature representations from large-scale pedestrians as references to compensate the missing details of small-scale pedestrians in feature space. Second, by assigning larger loss weights to small-scale instances, the scale-based focal loss can improve the performance on small-scale pedestrians but sacrifices the overall performance as shown in Table 6. Different from the scale-based focal loss, our SML improves both the performance on small-scale pedestrians and the overall performance. This is because our approach assigns the same weights to small-scale and large-scale pedestrians and can reduce the feature variance among small-scale and large-scale pedestrians.

Hyper-parameters. We experiment our method with different settings of hyper-parameters of H_S, H_L, and α, and the results can be found in supplementary material. We observe that our method is relatively stable under different hyper-parameter settings.

6.3 Results on Caltech

1× **Training Set.** Table 7 shows the results of the baseline and our approach with different network backbones, ResNet-18 and ResNet-50. Our approach improves the performance over the baseline, achieving 1.3% and 1.9% improvements with ResNet-18 and ResNet-50 respectively on the Medium subset, and 1.2% and 1.6% improvements with ResNet-18 and ResNet-50 respectively on the Far subset. The large improvements on tiny pedestrians demonstrate the effectiveness of SML on the small-scale pedestrians.

10× **Training Set.** In Table 8, we compare SML with the state-of-the-art methods, including RPN+BF[36], RepLoss [32], ALFNet [21], CSP [22] and OR-CNN [39]. To our best knowledge, all these methods in Table 8 use the new annotations for training and evaluation. Note that the ALFNet and CSP use data augmentation for training, e.g., random color distortion, horizontal flip and random crop, while SML is trained without any data augmentation techniques. Without pre-training on the CityPersons dataset, SML achieves the state-of-the-art performance on the Reasonable subset at the IoU thresholds of 0.5, and outperforms other the state-of-the-art methods on the

Table 8: Comparisons with state-of-the-art methods on the Caltech testing set. 10× **training set is used for training. + City means that the network is pre-trained on CityPersons.**

Method	Pre-train	Backbone	Reasonable		Small	
			IoU=0.5	IoU=0.75	IoU=0.5	IoU=0.75
RPN+BF[36]		VGG16	7.3	59.9	8.6	65.6
ALFNet [21]		ResNet-50	6.2	24.4	7.9	26.7
RepLoss [32]		ResNet-50	4.9	25.4	5.2	28.7
CSP [22]		ResNet-50	**4.5**	28.9	5.3	31.2
Baseline		ResNet-50	5.5	22.9	6.2	24.0
SML		ResNet-50	**4.5**	**21.3**	**4.6**	**23.6**
ALFNet [21] + City	✓	ResNet-50	4.5	19.4	6.2	23.6
RepLoss [32] + City	✓	ResNet-50	4.0	23.2	4.2	27.5
OR-CNN [39] + City	✓	VGG16	4.1	22.8	4.3	25.5
CSP [22] + City	✓	ResNet-50	3.8	22.0	4.7	24.8
SML+ City	✓	ResNet-50	**3.7**	**13.4**	**2.8**	**14.6**

Table 9: Comparisons with the baseline detector on the CityPersons validation set.

Backbone	Scale	Method	Reasonable	Small
ResNet-18	×1	Baseline	16.9	28.9
		SML	**15.9**	**26.1**
ResNet-50	×1	Baseline	13.4	21.9
		SML	**12.3**	**19.3**
ResNet-50	×1.3	Baseline	12.0	18.3
		SML	**10.6**	**16.3**

Reasonable subset at the IoU thresholds of 0.75 and Small subset at the IoU thresholds of 0.5 and 0.75, respectively. Notably, our proposed SML pre-trained on CityPersons (SML + City) achieves 3.7% MR^{-2} and 13.4% MR^{-2} on the Reasonable subset and 2.8% MR^{-2} and 14.6% MR^{-2} on the Small subset at the IoU thresholds of 0.5 and 0.75, achieving the best performance on the Caltech dataset. The qualitative examples on the Caltech dataset can be found in the supplementary material.

6.4 Results on CityPersons

Table 9 shows the results of the baseline and SML with different backbone networks. We can see that our approach outperforms the baseline detector on the Reasonable subset by 1.0% ∼ 1.4% in MR^{-2} with different backbone networks. Furthermore, SML significantly reduces the miss rate by 2.6% and 2.0% respectively using ResNet-50 with ×1 and ×1.3 input image scales and reduces the miss rate by 2.8% using ResNet-18 with ×1 input image scale on the Small subset. Qualitative results of our approach on the CityPersons dataset can be found in the supplementary material.

In Table 10, we compare our approach with state-of-the-art methods including Adapted Faster RCNN [38], TTL(MRF) [31], RepLoss [32], ALFNet [21], CSP [22], PDOE+RPN [44], OR-CNN [39], Adaptive NMS [20], and MGAN [27]. Besides, we also list the results of one-stage detector CSP with SML, which will be discussed in supplementary material. The experimental results in Table 10 shows that the proposed SML achieves comparable performance on the large-scale pedestrians and state-of-the-art miss rates on the small-scale pedestrians.

Table 10: Comparisons with state-of-the-art methods on the CityPersons validation set. The top 3 results are highlighted in red, blue and green, respectively. The CSP, ALFNet and CSP+SML use the original image scale.

Method	Backbone	Reasonable	Partial	Bare	Small
TTL [31]	ResNet-50	15.5	17.2	10.0	-
TTL+MRF [31]	ResNet-50	14.4	15.9	9.2	-
AF RCNN [38]	VGG16	12.8	-	-	-
ALFNet [21]	ResNet-50	12.0	11.4	8.4	19.0
RepLoss [32]	ResNet-50	11.6	14.8	7.0	-
PDOE+RPN [44]	VGG16	11.2	-	-	-
CSP [22]	ResNet-50	11.0	10.4	7.3	16.0
OR-CNN [39]	VGG16	11.0	13.7	5.9	-
Liu *et al* [20]	VGG16	10.8	11.4	6.2	-
MGAN [27]	VGG16	10.5	-	-	-
Baseline+SML	ResNet-50	10.6	10.1	6.4	16.3
CSP [22] +SML	ResNet-50	10.6	9.6	7.0	14.1

7 CONCLUSION

In this paper, we analyze the low recall and limited detection performance of small-scale pedestrian and reveal that the main cause is misclassification of small instances. Based on the analysis, we propose a Self-Mimic Learning method to enhance the representations for small-scale pedestrians and reduce intra-class feature variance by mimicking rich representations from large-scale pedestrians. To achieve this, we enforce the feature representations of small-scale pedestrians to approach those of large-scale pedestrians by making the RoI features of small-scale pedestrians mimic the local average RoI features of large-scale pedestrians. Our approach is a general component which can be efficiently applied to both one-stage and two-stage detectors with any backbone network to improve the feature representations of small-scale pedestrians. Exhaustive experiments on both Caltech and CityPersons datasets validate the effectiveness and superiority of our approach.

Acknowledgment. This work is supported in part by the start-up funds from State University of New York at Buffalo and gift grant from Horizon Robotics.

REFERENCES

[1] Lei Jimmy Ba and Rich Caruanaa. 2014. Do deep nets really need to be deep. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

[2] Garrick Brazil and Xiaoming Liu. 2019. Pedestrian detection with autoregressive network phases. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7231–7240.

[3] Jiale Cao, Yanwei Pang, Jungong Han, Bolin Gao, and Xuelong Li. 2019. Taking a look at small-scale pedestrians and occluded pedestrians. *IEEE Transactions on Image Processing*, 3143–3152.

[4] Jiale Cao, Yanwei Pang, and Xuelong Li. 2016. Pedestrian detection inspired by appearance constancy and shape symmetry. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1316–1324.

[5] Jiale Cao, Yanwei Pang, and Xuelong Li. 2017. Learning multilayer channel features for pedestrian detection. *IEEE Transactions on Image Processing*, 3210–3220.

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 764–773.

[7] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. 2012. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 743–761.

[8] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 1440–1448.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 580–587.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 2961–2969.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[13] Xin Huang, Zheng Ge, Zequn Jie, and Osamu Yoshie. 2020. NMS by Representative Region: Towards Crowded Pedestrian Detection by Proposal Pairing. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10750–10759.

[14] Yonghyun Kim, Bong-Nam Kang, and Daijin Kim. 2018. SAN: Learning relationship between convolutional features for multi-scale object detection. In *Proceedings of European conference on computer vision (ECCV)*. 316–331.

[15] Hongyang Li, Bo Dai, Shaoshuai Shi, Wanli Ouyang, and Xiaogang Wang. 2019. Feature intertwiner for object detection. In *Proceedings of International Conference on Learning Representations (ICLR)*.

[16] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. 2017. Perceptual generative adversarial networks for small object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1222–1230.

[17] Quanquan Li, Shengying Jin, and junjie Yan. 2017. Mimicking very efficient network for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6356–6364.

[18] Chunze Lin, Jiwen Lu, Gang Wang, and Jie Zhou. 2018. Graininess-aware deep feature learning for pedestrian detection. In *Proceedings of European conference on computer vision (ECCV)*. 732–747.

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2117–2125.

[20] Songtao Liu, Di Huang, and Yunhong Wang. 2019. Adaptive nms: refining pedestrian detection in a crowd. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6459–6468.

[21] Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. 2018. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of European conference on computer vision (ECCV)*. 618–634.

[22] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. 2019. High-level semantic feature detection: a new perspective for pedestrian detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5187–5196.

[23] Yan Luo, Chongyang Zhang, Muming Zhao, Hao Zhou, and Jun Sun. 2020. Where, What, Whether: Multi-modal learning meets pedestrian detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 14065–14073.

[24] Junhyug Noh, Soochan Lee, Beomsu Kim, and Gunhee Kim. 2018. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 966–974.

[25] Wanli Ouyang and Xiaogang Wang. [n.d.]. Joint deep learning for pedestrian detection.

[26] Yanwei Pang, Jiale Cao, Jian Wang, and Jungong Han. 2019. LJCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images. *IEEE Transactions on Information Forensics and Security*, 3322–3331.

[27] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. 2019. Mask-Guided attention network for occluded pedestrian detection. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 4967–4975.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal network. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

[29] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *Proceedings of International Conference on Learning Representations (ICLR)*.

[30] Bharat Singh and Larry S. Davis. 2018. An analysis of scale invariance in object detection – SNIP. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3578–3587.

[31] 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. In *Proceedings of European conference on computer vision (ECCV)*. 536–551.

[32] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. 2018. Repulsion loss: Detecting pedestrian in a crowd. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7774–7783.

[33] Yi Wei, Xinyu Pan, Hongwei Qin, Wanli Ouyang, and Junjie Yan. 2018. Quantization mimic: Towards very tiny cnn for object detection. In *Proceedings of European conference on computer vision (ECCV)*. 267–283.

[34] Jialian Wu, Liangchen Song, Tiancai Wang, Qian Zhang, and Junsong Yuan. 2020. Forest R-CNN: Large-vocabulary long-tailed object detection and instance segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia (MM)*.

[35] Jialian Wu, Chunluan Zhou, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. 2020. Temporal-context enhanced detection of heavily occluded pedestrians. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 13430–13439.

[36] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. 2016. Is faster r-cnn doing well for pedestrian detection. In *Proceedings of European conference on computer vision (ECCV)*. 443–457.

[37] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. 2016. How far are we from solving pedestrian detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1259–1267.

[38] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. 2017. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3213–3221.

[39] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. 2018. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In *Proceedings of European conference on computer vision (ECCV)*. 637–653.

[40] Shanshan Zhang, Jian Yang, and Bernt Schiele. 2018. Occluded pedestrian detection through guided attention in cnns. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6995–7003.

[41] Chunluan Zhou, Ming Yang, and Junsong Yuan. 2019. Discriminative feature transformation for occluded pedestrian detection. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 9557–9566.

[42] Chunluan Zhou and Junsong Yuan. 2016. Learning to integrate occlusionspecific detectors for heavily occluded pedestrian detection. In *Proceedings of Asian Conference on Computer Vision (ACCV)*. 305–320.

[43] Chunluan Zhou and Junsong Yuan. 2017. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 3486–3495.

[44] Chunluan Zhou and Junsong Yuan. 2018. Bi-box regression for pedestrian detection and occlusion estimation. In *Proceedings of European conference on computer vision (ECCV)*. 135–151.

[45] Chunluan Zhou and Junsong Yuan. 2019. Multi-label learning of part detectors for occluded pedestrian detection. *Pattern Recognition*, 99–111.

[46] Chunluan Zhou and Junsong Yuan. 2020. Occlusion pattern discovery for partially occluded object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2067–2080.

[47] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. 2019. Deformable convNets v2: more deformable, better results. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 9308–9316.