Sequential Attention-Based Distinct Part Modeling for Balanced Pedestrian Detection

Yan Luo, Chongyang Zhang, *Member, IEEE*, Weiyao Lin, *Senior Member, IEEE*, Xiaokang Yang, *Fellow, IEEE*, and Jun Sun, *Member, IEEE*

***Abstract*— Despite pedestrian detectors having made significant progress by introducing convolutional neural networks, their performance still suffers degradation, especially in occlusion scenes with more false positives (FPs) and false negatives (FNs). To alleviate the problem, we propose a novel Sequential Attention-based Distinct Part Modeling (SA-DPM) for balanced pedestrian detection. It takes one step further in constructing more robust representation that supports detection with fewer FNs and FPs. Specifically, the Sequential Attention serves as one internal perception process that captures several distinct part areas step by step from each pedestrian proposal (full-body). Different from the previous either-or feature selection, the following Joint Learning attempts to seek a reasonable trade-off between part and full-body features, and combines both features for more accurate classification and regression. Evaluation on the widely used pedestrian datasets including Caltech and Citypersons shows that the proposed SA-DPM achieves promising performance for both non-occluded and occluded pedestrian detection tasks, especially on Caltech Heavy Occlusion set, which yields a new state-of-the-art miss rate by 30.18% and outperforms the second best detector by 6.32%.**

***Index Terms*— Pedestrian detection, occlusion handling, part attention.**

I. Introduction

PEDESTRIAN detection is highly valued in many applications [1], such as autonomous driving, in which accurate and robust pedestrian detection results have a direct impact on the planning and decision of autonomous vehicles. In recent years, many works have been devoted to this detection task [2]–[6]. Despite that many recent detectors work reasonably well with pedestrians under simple scenarios, their performance always sustains a significant deterioration in the presence of occlusion.

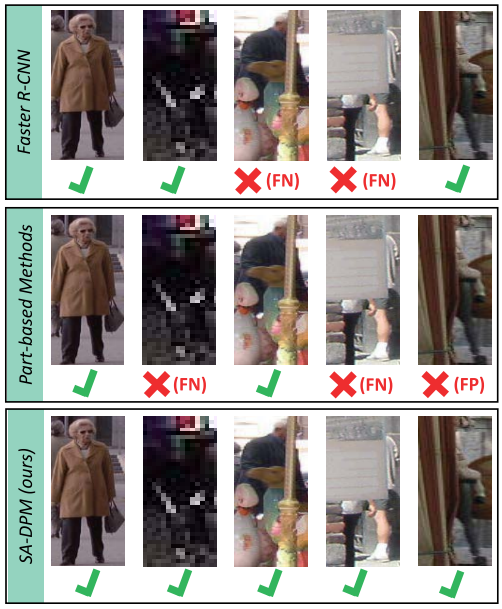


Fig. 1. Illustration of different detectors. Methods such as Faster R-CNN tend to have better performance on full-body (non-occluded) pedestrian detection, but show failure on occluded instances. Some part-based methods could effectively deal with some occlusion issues, but they still face the problems of false negatives with small-scale pedestrians and increasing false positives. Ours performance shows in the third row with less false negatives and false positives. **FN** means false negatives and **FP** means false positives.

Therefore, it is essential to seek an effective method to solve the occluded pedestrian detection problem. A simple and straightforward method proposed by researchers in the past few years is to detect or perceive the visible pedestrian part firstly. With the well-trained detector, these methods utilize the visible part features to locate each occluded pedestrian and thus effectively reduce the false negatives (FNs). Typical practice falls into above paradigm including [3], [4], [7], [8].

Even though aforementioned detectors succeed in identifying occluded pedestrian instances, there still exist two main challenges. *(1) False Positives.* Part features have been shown to be discriminative, but at the same time, they also tend to be biased. Therefore, when it comes to suspicious objects that have parts similar to pedestrians (e.g., trunks *vs.* body, car lights *vs.* head), the over-preference on part features usually brings about more false positives. For example, in the last column of Figure 1, one negative instance with similar lowerbody appearance, is taken as positive pedestrian falsely. (2) *False Negatives.* These part-based methods highly depend

Manuscript received 19 April 2021; revised 12 October 2021 and 16 December 2021; accepted 31 December 2021. Date of publication 26 January 2022; date of current version 12 September 2022. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61971281, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102, and in part by the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant 18DZ2270700. The Associate Editor for this article was H. Dong. *(Corresponding author: Chongyang Zhang.)*

Yan Luo, Weiyao Lin, and Jun Sun are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: luoyan_bb@sjtu.edu.cn; wylin@sjtu.edu.cn; junsun@sjtu.edu.cn).

Chongyang Zhang and Xiaokang Yang are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the MoE Key Laboratory of Artificial Intelligence, Artificial Intelligence Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: sunny_zhang@sjtu.edu.cn; xkyang@sjtu.edu.cn).

Digital Object Identifier 10.1109/TITS.2022.3144359

on the accuracy of the part detector. It means the inaccurate part detection could directly affect the overall performance. Especially for small-scale pedestrians with blurred boundaries and obscure appearance, it is difficult to detect body parts accurately. As is shown in the middle column of Figure 1, the pedestrian target that is not occluded but has obscure boundaries, is detected as a negative sample incorrectly.

The above challenges make the current pedestrian detectors trapped in the unbalanced performance. Over-preference on part features could lead to the increase of false positives (FPs) on non-occluded pedestrians, while full-body detectors bring in false negatives (FNs) of occluded ones. To solve this problem, we propose a novel Sequential Attention-based Distinct Part Modeling (SA-DPM) that combines both part and full-body features for more balanced pedestrian detection. Specifically, since part features have shown to be capable of providing fine-grained information for occlusion handling, we propose the sequential attention to perceive the distinct part areas. Different from previous methods that introduce extra part annotations to train one better part detector, the proposed module is optimized with the weakly supervised method. It decouples the detection process into channel and spatial dimensions for dynamic part perception. The sequential manner means the perception process is step by step, and each step consisted of both channel and spatial attention is to capture its most distinct part area. Besides, by the constraints of *Frobenius* norm, the part areas from the different steps are as different as possible. It guarantees the diversity of the parts we extract. In this way, the network is promoted to focus on several relevant and visible part areas to establish robust feature representation against occlusion issues, so that false negatives under occlusion scenes could be largely alleviated.

As discussed above, the performance with the single part features is far from being satisfactory. Over-weighting on parts usually brings about FPs on general scenes and FNs on small-scale instances that degrade the overall performance. Therefore, we develop a joint representation learning to seek a reasonable trade-off between part and full-body features. It means multi-components are re-weighted by corresponding weighting coefficients. In this way, the network performs feature re-calibration, and thus it can adaptively emphasize informative features and suppress less useful ones under different circumstances. Furthermore, in addition to generating more robust features of each pedestrian instance, we also consider further improve performance during the post-processing process. The heavy occlusion between pedestrians imposes great challenges to the standard Non-Maximum Suppression (NMS). The IoU threshold is difficult to set in the occlusion scenario. A relatively low threshold tends to perform better in sparse pedestrian scenes, but can lead to missed detection in dense pedestrian scenes. To avoid such a dilemma, we propose the Balanced Non-Maximum Suppression (BNMS) leveraging both the part and full-body boxes to effectively refine the detection results.

A brief flow chart is shown in Figure 2. Generally speaking, in our proposed SA-DPM (the lower row in Figure 2), the proposals are first fed into Sequential Attention (**SA**) to extract part features. Next the part and full-body features are

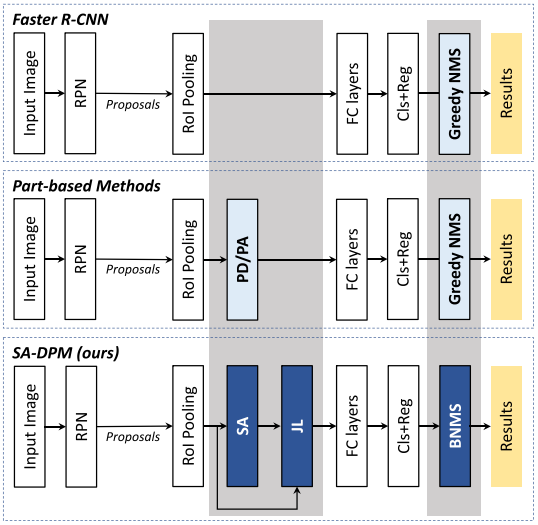


Fig. 2. Comparison of different pedestrian detection paradigms. "**PD/PA**" means the Part Detector/Part Attention mechanism. "**SA**" means the Sequential Attention. "**JL**" means the Joint Learning module. "**BNMS**" means the post-processing mechanism of Balanced Non-Maximum Suppression. The upper one is the Faster R-CNN, which has been introduced in many pedestrian detectors but show unsatisfactory performance on occluded instances. The middle one is part-based methods with part attention mechanism or extra part detectors. The lower one is our proposed SA-DPM, which is inspired by the above two, but more accurate.

sent to Joint Learning mechanism (**JL**) to learn weighting parameters. Furthermore, based on the estimated part and full-body boxes, we introduce the Balanced Non-Maximum Suppression (**BNMS**) to retain more high-quality detection results, especially for occluded pedestrian detection task.

To the best of our knowledge, this is the first attempt to introduce sequential attention for pedestrian detection. Without bells and whistles, empirical evaluation reveals that the novel SA-DPM leads to new state-of-the-art performance on several widely-used datasets including Caltech [9], Citypersons [10]. For instance on Caltech, SA-DPM achieves 3.45% MR^{-2} on reasonable subset. Besides, compared with other state-of-the-art methods, we have achieved more robust performance not only on regular datasets, but also on challenging settings, such as occlusion subsets. In summary, our key contributions are as follows:

- It is the first to put forward the sequential attention for pedestrian detection by learning not only where to focus on but also how to aggregate features of part and full-body regions to ensure fewer FPs and FNs at the same time.
- Under the constraints of Frobenius norm, the spatial and channel attention attempt to sequentially extract several distinct parts to alleviate FNs in occluded scenes. Moreover, the proposed network is trained using a weakly supervised method. It enables efficient part localization without introducing an elaborate part detector with expensive body annotations.
- We attempt to seek the reasonable trade-off between full-body and part features with a set of normalized weights and develop one Balanced NMS. It makes full use of the perceived part areas and effectively refines the detection results.

The rest of this paper is organized as follows. Section II presents a brief introduction of related works. Section III

elaborates the detailed structure of the proposed network. Next, we provide experimental results and discussion in Section IV. Finally, the paper is concluded in Section V.

II. Related Work

A. Pedestrian Detection With CNNs

Because of its great value in automatic driving and video surveillance, pedestrian detection has attracted wide attention over the past decade. Generally speaking, pedestrian detection serves as the specific task and its mainstream frameworks still follow the basic practice in general object detection. It can be summarized into two categories. One of the categories is anchor-based methods. Cai *et al.* [11] proposed a network combined with a proposal sub-network and a detection sub-network, so that receptive fields can match objects of different scales. Du *et al.* [12] proposed a novel network fusion method called soft-rejection based network fusion. It employed a classification network, consisting of multiple deep neural network classifiers, to refine the pedestrian candidates. In addition, there are also methods focusing on occluded pedestrian detection. Wang *et al.* [2] proposed a novel bounding box regression loss specifically designed for crowd scenes, termed as repulsion loss. Another typical detector called CSP [13] can be classified as the other category, namely anchor-free methods. CSP discards the preset of the fixed-size anchors, but predicts the confidence of each point as the pedestrian center instead. It is no doubt that these methods have achieved great success in general pedestrian detection. However, when facing crowded scenes, most detectors have relatively poor performance with more FPs and FNs for lacking of mechanisms oriented towards occluded instances.

B. Pedestrian Detection Based on Part Features

Pedestrian detection based on part features is one of effective methods in crowded scenes. Part features have shown to be capable of providing important local cues of human appearance. It is natural to extract part features to further improve pedestrian detectors in some state-of-the-art methods, such as [3], [4]. DPM (Deformable Part Model) [14] sets the artificial constraint on pedestrian parts and trains several classifiers to identify different part areas according to HOG features (Histogram of Oriented Gradient). As pedestrians usually form a class with high intra-class similarity, some detectors [4], [15] also employ a statistical model of the upright human body where the head, the upper-body and the lower-body are strictly restricted in regions with specific ratio. This method is simple and effective for pedestrian detection, but the hard constraint on part regions limits its performance in general scenes. Besides, Bi-box [16] divides pedestrian detection into two branches. One branch is used for full-body pedestrian detection, and the other branch is used for occluded pedestrian detection. Some other detectors [17] attempt to introduce extra part annotations to train a part detector, e.g. head detector. Though these works get promising performance in pedestrian detection, these approaches are somewhat costly by using supernumerary supervision. Different from above methods, our proposed self-guided perception brings unique advantages that regardless of a wide range of occlusion patterns, our framework is self-driven to place emphasis on relevant part features without introducing extra label information.

C. Attention Mechanism in Pedestrian Detection

Attention mechanism has been widely used in many fields such as Nature Language Processing (NLP) [18], image classification [19] and so on. Vaswani *et al.* [18] is the first attempt to introduce the self-attention mechanism to associate information from different locations in the input sequence and then calculate expression of entire sequence. Jie *et al.* [19] improved the expressive ability of network by accurately modeling the interaction relationship between channels of convolution feature. Zhang *et al.* [3] is the first to analyze channel-wise attention for pedestrian detection. It proposed channel-wise mechanisms to learn proper attention parameters for different channels so as to handle different occlusion patterns effectively. Pang *et al.* [8] attempts to introduce the mask-guided attention that promotes the network to focus on visible part features instead of the occluded ones.

Our approach is motivated by the success of attention modules in the above works. On one hand, we introduce the sequential attention. It consists of spatial and channel attention to sequentially extract distinct part features to alleviate FNs in crowded scenes. On the other hand, the following joint learning attempts to seek a reasonable feature aggregation method to keep fewer FNs and FPs simultaneously.

III. Proposed Method

A. Overview of the Proposed Method

As illustrated in Figure 1, the proposed model is based on the Faster R-CNN [20], with which proposals will be generated by RPN (Region Proposal Network). Each pooled proposal is then fed into the Sequential Attention module to predict K ($K = 3$ in this work) most distinct part areas ($\tilde{\mathbf{X}}^{(1)}$, $\tilde{\mathbf{X}}^{(2)}$, $\tilde{\mathbf{X}}^{(3)}$). The following joint learning adaptively weights on both full-body and part features to generate more robust representation for classification and regression. In the post-processing process, Balanced NMS is utilized to effectively refine the predicted boxes considering both full-body and part information estimated by above modules. More details of the proposed model are illustrated in Figure 3 and are described in the following subsections.

B. Sequential Attention-Based Distinct Part Modeling

In order to reduce false negatives in occlusion scenes, this module is proposed to capture distinct part features. It performs as sequentially cascaded steps, and each step consisted of both channel and spatial attention to capture the most relevant body part regions. More details can be seen in Figure 4. We take the three-step manner as an example.

1) Sequential Channel Attention: Sequential Channel Attention plays a key role in capturing *distinct* body parts. Specifically, its design is based on the following motivations:

- Considering that different channels could activate responses for different body parts [3], the proposed channel

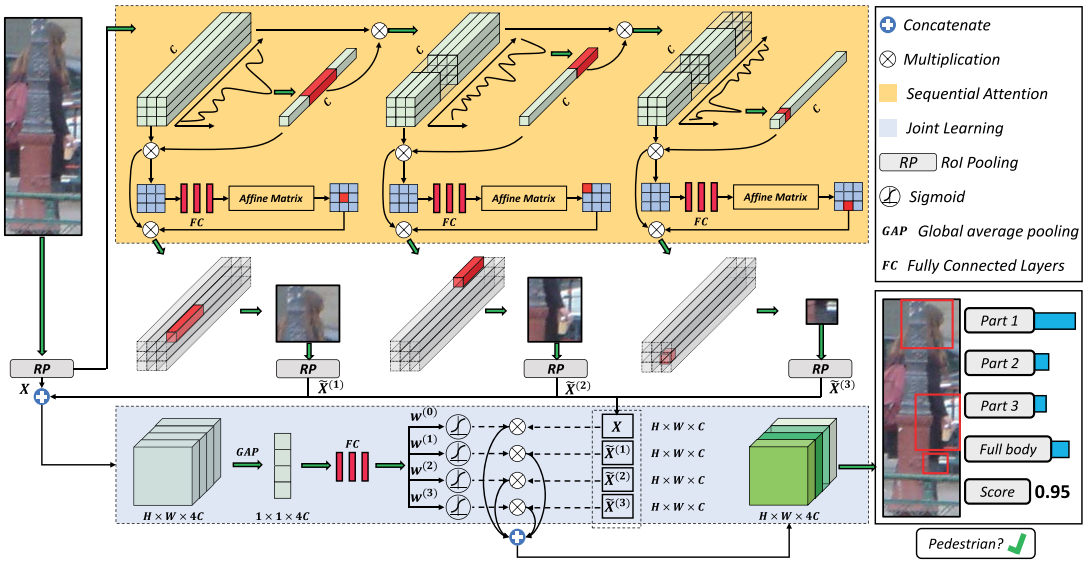


Fig. 3. Proposed SA-DPM for pedestrian detection. It takes the proposal as input and outputs the detection results. The input proposal is first fed into the Sequential Attention that outputs several distinct part areas. The following Joint Learning combines both part and full-body features for more accurate detection results.

attention can be regarded as the selection of various semantic part information on the demand of input pedestrian features.

• Some of the previous methods could be more biased towards a specific part feature, such as head. We attempt to capture the most discriminative part areas rather than the fixed one in each proposal. Therefore, each step predicts the strongest response *interval* across channels, and in order to guarantee the diversity of the part areas, *Frobenius* norm is also introduced to keep linear interdependence between parts in the different steps.

Let $\mathbf{X} \in \mathbb{R}^{C\times W\times H}$ be the pooled feature of one proposal, where H and W are the spatial height and width, and C is the number of channels. The proposed channel attention first predicts the strongest part response across channels. To this end, two convolution kernels are introduced to calculate where the strongest response is and what the corresponding interval is, formulated as:

$$p_m^{(1)} = \sum_{i=m-[\frac{H}{2}]}^{i=m+[\frac{H}{2}]} \sum_{j=1}^{W} \sum_{k=1}^{H} \mathbf{X}_{i,j,k} \cdot \mathbf{K}_{i+1-m+[\frac{H}{2}],j,k}$$
$$\sigma_m^{(1)} = \sum_{i=m-[\frac{H}{2}]}^{i=m+[\frac{H}{2}]} \sum_{j=1}^{W} \sum_{k=1}^{H} \mathbf{X}_{i,j,k} \cdot \mathbf{Q}_{i+1-m+[\frac{H}{2}],j,k} \quad (1)$$

where $\mathbf{K}$ and $\mathbf{Q}$ represent the parameters of convolution kernel to predict the probability and variance, respectively. H and W are the corresponding height and width of the pooled feature, and are both set to 7 in our experiments. $p_m^{(1)}$ and $\sigma_m^{(1)}$ represent the probability of the m-th channel to be the center of the part region and the variance to calculate the corresponding channel intervals in the first step. Considering all the channels, the probability vector of the first step can be described as:

$$\mathbf{p}^{(1)} = [p_1^{(1)}, p_2^{(1)}, p_3^{(1)}, \ldots, p_C^{(1)}]$$
$$\tilde{\mathbf{p}}^{(1)} = \text{softmax } \mathbf{p}^{(1)} \quad (2)$$

The largest element in the vector $\tilde{\mathbf{p}}^{(1)}$ is the center of the channel interval. Besides, with the estimated variance $\sigma_m^{(1)}$, the corresponding interval length ϵ can be calculated via Gaussian modeling as the following:

$$max\ \epsilon$$
$$\text{s.t.} \int_{-\epsilon/C}^{\epsilon/C} \frac{1}{\sqrt{2\pi}\,\sigma_m^{(n)}} \left\{ exp(-\frac{x^2}{2[\sigma_m^{(n)}]^2}) \right\} dx < \delta_0 \quad (3)$$

where δ_0 is the hyper-parameter to control the uncertainty of the part interval, and sets as 0.8 in our experiments. It is a difficult thing to pick up Eq 3 directly. So we just do an estimation of the value ϵ, using an enumeration manner. On top of above, the channel interval that represents the extracted part features in the first step can be defined as $[\tilde{m}-\epsilon, \tilde{m}+\epsilon]$, in which $\tilde{m}$ represents the $\tilde{m}$-th channel with the maximum probability of $\tilde{\mathbf{p}}^{(1)}$ and ϵ is estimated in Equation 3. The total $2\epsilon + 1$ channels are mapped back to the original proposal, and the obtained strongest response region is then fed into RoI pooling to generate the channel part feature $\mathbf{X}^{(1)}$ of the first step.

Specifically, for other steps except the first one, we have made the modifications to the input features, in which we expect the previous strong response channels will be suppressed under the current step, and thus enable the framework to focus more on distinct part regions. In order to describe the process more clearly, we first define the channel attention process of the first step with the input feature $\mathbf{X}$ as:

$$\mathbf{X}^{(1)} = \varphi(\mathbf{X}) \quad (4)$$

where $\varphi(\cdot)$ represents the parameters to optimize. Considering the following steps, the outputs can be denoted as:

$$\mathbf{X}^{(n)} = \varphi[\prod_{q=1}^{n-1} \text{diag } (1 - \tilde{\mathbf{p}}^{(q)}) \cdot \mathbf{X}], \quad n = 2, 3 \cdots \quad (5)$$

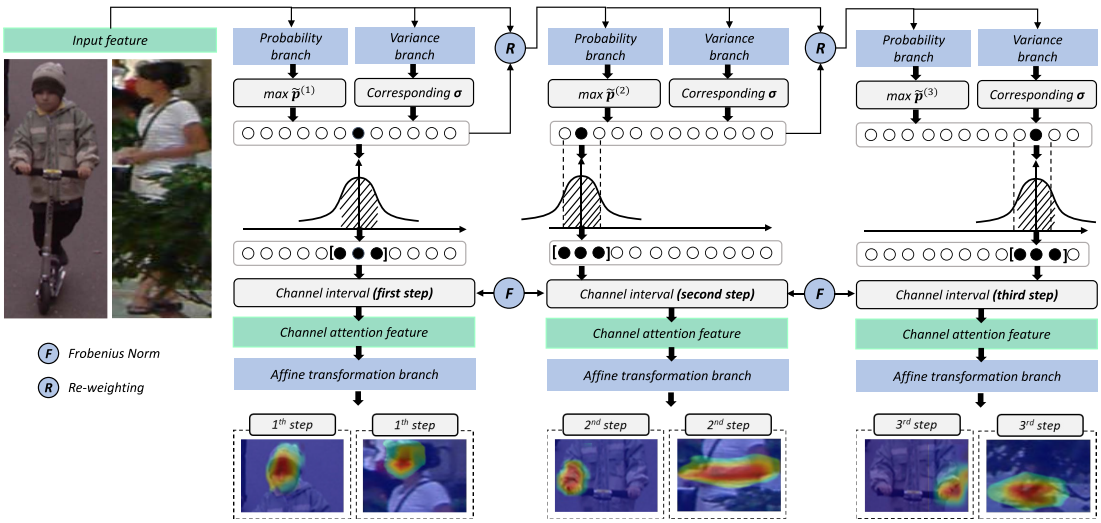


Fig. 4. Details on the proposed Sequential Attention. The upper row shows the process of channel attention, and the lower one represents the spatial attention. The diagrammatic sketch of input and output results shows that distinct part regions can be extracted via the proposed mechanism.

where $\tilde{\mathbf{p}}^{(q)}$ represents the probability vector of the former q-th step and has been clearly defined in Equation 2. Besides, n means the maximum number of steps. We take the $n = 3$ for an example, which means the sequential attention contains three steps. And the diag means the transformation that promotes the vector $\tilde{\mathbf{p}}$ to the corresponding diagonal matrix. In order to further ensure that distinct body parts are extracted across different steps, *Frobenius* norm is also introduced to measure the degree of linear independence between different part features, denoted as:

$$\begin{aligned} \mathbf{P}_l &= [\tilde{\mathbf{p}}^{(1)}; \tilde{\mathbf{p}}^{(2)}; \ldots; \tilde{\mathbf{p}}^{(l)}] \\ \mathcal{L}_F &= \sum_{l=1}^{n} \left\{ ||\sqrt{\mathbf{P}_l \mathbf{P}_l^T}||_F^2 - \sum_{u=1}^{l} [\tilde{\mathbf{p}}^{(u)}]^2 \right\} \end{aligned} \tag{6}$$

where $||\cdot||_F$ represents the *Frobenius* norm of the corresponding matrix, l means the l-th step and $\mathcal{L}_F$ is the corresponding loss to estimate the degree of linear independence of all steps. If $\mathcal{L}_F$ is zero, the part features extracted by various steps are completely linear independent.

2) Spatial Attention: As discussed above, various occlusion patterns are the key reason that restricts the detection performance of occluded pedestrians. Different from attempting to model such a large number of occlusion patterns, the proposed method tries to map the visible part feature into a manifold of much lower dimensionality than that of the original occlusion space. In other words, the changes of visible part regions are constrained in a low dimensional space, which is defined as *affine transformation* in the proposed method that maps the original part regions into the three degrees of freedom of variability including translations, scalings and rotations. Note that the transformation parameters are served as latent variables, which are embedded by the input proposal features and used to reconstruct accurate part features.

Specifically, we consider the output features of Channel Attention of n-th step $\mathbf{X}^{(n)} \in \mathbb{R}^{\mathrm{C} \times \mathrm{W} \times \mathrm{H}}$, where H and W are the spatial height and width, and C is the number of channels. In general, spatial attention performs the following transformation:

$$\begin{aligned} \kappa^{(n)} &= \gamma(\mathbf{X}^{(n)}) \\ \tilde{\mathbf{X}}^{(n)} &= \mathcal{M}(\mathbf{X}^{(n)}, \kappa^{(n)}) \end{aligned} \tag{7}$$

where $\mathbf{X}^{(n)}$ is the output of the n-th step, $\gamma(\cdot)$ is the parameters to optimize, $\kappa^{(n)}$ performs as the latent variable, and $\mathcal{M}(\cdot)$ re-encodes the input feature $\mathbf{X}^{(n)}$ based on the estimated transformation parameters $\kappa^{(n)}$. In details, the proposed affine transformation is the fusion of three degrees of freedom of variability including the translation factors (t_x, t_y), the rotation factor θ, and the scaling factors (s_x, s_y). The five parameters are the outputs of three fully-connected layers, and consists of the latent variable $\kappa^{(n)}$. Given $\kappa^{(n)}$, the transformation matrix $\mathbf{T}$ is as follows:

$$\begin{aligned} \mathbf{T} &= \mathbf{I_1} \mathbf{I_2} \mathbf{I_3} \\ &= \begin{bmatrix} s_x cos\ \theta & s_y sin\ \theta & t_x \\ -s_x sin\ \theta & s_y cos\ \theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{8}$$

where $\mathbf{I_1}$, $\mathbf{I_2}$ and $\mathbf{I_3}$ are the standard affine transformation matrix corresponding to translations, rotations and scalings, respectively. The transformation process maps all 2D coordinates in the input $\mathbf{X}^{(n)}$. The original horizontal part region is then transformed into a reduced polygonal box and the final output of the Sequential Attention is sequentially predicted, denoted as $\tilde{\mathbf{X}}^{(n)}$.

As illustrated in Figure4, to summarize the above sequential manner, the input proposal feature $\mathbf{X}$ is first extracted by channel attention and outputs various channel part features $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$. The channel feature of each step is then fed into the spatial attention module, respectively, and outputs first step feature denoted as $\tilde{\mathbf{X}}^{(1)}$, $\tilde{\mathbf{X}}^{(2)}$ and $\tilde{\mathbf{X}}^{(3)}$.

C. Joint Part and Full-Body Learning

Only using the feature of full-body could result in more false negatives, while the single part feature could bring about false positives. The purpose of Joint Learning is to seek an

optimal trade-off between part and full-body features, which attempts to selectively emphasize informative features and suppress less useful ones directing against various situations. Two mechanisms are adopted in the proposed Joint Learning. The first one aims to re-weight both part and full-body features to generate a reasonable representation for downstream pedestrian classification and localization. The second one presents a simple post-processing method to refine the detection results based on the pre-predicted part and full-body regions.

The full-body and part features are first concatenated along channels, denoted as $\mathbf{D} \in \mathbb{R}^{4C\times H\times W}$. Due to the mixed feature $\mathbf{D}$ entangled with the local spatial correlation, the network contains Global Average Pooling (GAP) to describe more specific component information. Estimated by the fully-connected layers, importance of different components can be represented as follows:

$$w^{(\mathrm{n})} = \psi[\mathcal{G}(\mathbf{D})] \tag{9}$$

where $\psi(\cdot)$ is the function of fully-connected layers to optimize, $\mathcal{G}(\cdot)$ is the GAP function, and $w^{(n)}$ is the value to measure feature importance of different steps. When n is zero, $w^{(0)}$ represents the contribution of the full-body feature. During training, we think each proposal should have at least one component responsible for its classification, and thus introduce a normalized loss, so that $\sum_{q=0}^{n} w^{(q)} \approx 1$. This additional constraint promotes the network to focus on different components at the same point and thus adaptively weight discriminative features. Therefore, the overall loss function is modeled as follows:

$$\mathcal{L} = \mathcal{L}_{cls}(b_r, g_r^*) + \mathcal{L}_{reg}(b_r, g_r^*) + \lambda_1[1 - \sum_{q=0}^{n} w^{(q)}]^2 + \lambda_2\mathcal{L}_F \tag{10}$$

where b_r is the r-th detected box being a pedestrian, g_r^* is the corresponding ground truth class, $\mathcal{L}_{cls}$ is the classification loss and $\mathcal{L}_{reg}$ is the regression loss.

Discussion: By design, the Sequential Attention attempts to perceive several distinct parts in each proposal to generate robust feature representation against occlusion issues. Specifically, the differences between the deep model in this paper and the previous methods [3], [14] are as follows:

- Compared with attention methods, e.g., Faster R-CNN+ATT [3], ours performs as a sequential manner that perceives several distinct part areas step by step. Besides, with the combination of Joint Learning module, ours not only learns part areas but also learns how to aggregate features of part and full-body regions. It ensures the balanced performance on both occluded and non-occluded pedestrian detection.
- Compared with DPM [14], ours pays more attention to the strongest response part in each step (Figure 4) rather than the pre-fixed part templates. In this way, the proposed part estimation is conditional more on the input features, and thus performs more flexibly under various occlusion scenes. Besides, ours shows more robust performance compared with head detector based methods [7], [21] for the same reason.

Fig. 5. Balanced Non-Maximum Suppression. If the previous NMS is used, without a suitable threshold, some boxes are easily suppressed, such as the red solid box on the left of the figure above. We aim to use the most discriminative part to distinguish the box. For example, on the right of the figure above, we use the head (red dashed box) instead of the full body to calculate the IoU in NMS.

D. Balanced Non-Maximum Suppression

In addition, we discover that previous Non-Maximum Suppression (NMS) with the full-body boxes will bring about more FNs with a lower NMS threshold, while a higher one may bring about more FPs. As illustrated in Figure 5 (left), the red solid box is easily suppressed by the green one under occlusion cases, in which the large overlap area often occurs between pedestrian full-body boxes. In this work, we naturally predict the full body boxes and part masks, which provides the possibility to improve previous NMS.

The proposed Balanced NMS leverages the distinct part boxes, which effectively averts the troubles brought by the Greedy NMS (GNMS) on highly overlapped full bodies. Detailed algorithm is shown in Algorithm 1. The part boxes denoted as B^1, B^2 and B^3 are generated from the most relevant part area (highlighted regions in Figure 4) with the minimal outer rectangles. Different from the Greedy NMS, which mainly uses full-body to measure whether the two boxes represent the same pedestrian, the IoU of both part and full-body are all taken into consideration in the proposed BNMS. Since the visible parts of pedestrians usually are free from occlusion issues, the IoU between the distinct part regions is a better indicator showing whether the box should be retained or suppressed. As illustrated in Figure 5, although the two pedestrians are occluded seriously, which brings about a dilemma for Greedy NMS to refine the detection results, the proposed BNMS with the IoU calculated between part boxes can still distinguish the two instances well. The core of the algorithm is that according to the weight calculated in Equation 9, in different scenarios, it will tend to use different parts to refine the predicted boxes. As a result, the introduced BNMS greatly reduces false negatives and false positives, and significantly improves performance of occluded instances.

IV. Experiments

We assess the effectiveness of our proposed method for pedestrian detection on widely used datasets Caltech [9], [28] and Citypersons [10].

Algorithm 1 Balanced Non-Maximum Suppression

Input:
B^0: detection boxes set of full-body.
B^1, B^2 and B^3: part set in the first, second and third step.
W: the weighting set of $w^{(0)}$, $w^{(1)}$, $w^{(2)}$ and $w^{(3)}$.
S: the score set of full-body boxes.
N_t: threshold of BNMS.
Output:
D: detection results.

$D \leftarrow \{\}$, $B \leftarrow B^0 \cup B^1 \cup B^2 \cup B^3$
while $B \neq empty$ **do**
  $\tilde{S} \leftarrow \max S$
  $\tilde{B}^0 = \left\{\tilde{b}^0\right\} \leftarrow \tilde{S}$, $\tilde{B}^1 = \left\{\tilde{b}_1\right\} \leftarrow \tilde{S}$, $\tilde{B}^2 = \left\{\tilde{b}_2\right\} \leftarrow \tilde{S}$, $\tilde{B}^3 = \left\{\tilde{b}_3\right\} \leftarrow \tilde{S}$
  $\tilde{B} \leftarrow \tilde{B}^0 \cup \tilde{B}^1 \cup \tilde{B}^2 \cup \tilde{B}^3$
  $D = D \cup \tilde{B}$, $B = B - \tilde{B}$, $S = S - \tilde{S}$
  for $b_r^0 \in B^0, b_r^1 \in B^1, b_r^2 \in B^2, b_r^3 \in B^3$ **do**
    $N_r = w_r^{(0)} \cdot \mathbf{IoU}(\tilde{b}^0, b_r^0) + w_r^{(1)} \cdot \mathbf{IoU}(\tilde{b}^1, b_r^1) + w_r^{(2)} \cdot \mathbf{IoU}(\tilde{b}^2, b_r^2) + w_r^{(3)} \cdot \mathbf{IoU}(\tilde{b}^3, b_r^3)$
    if $N_r \geq N_t$ **then**
      $B_r \leftarrow \{b_r^0, b_r^1, b_r^2, b_r^3\}$, $B = B - B_r$, $S = S - S_r$
    end if
  end for
end while
return D

A. Experimental Setup

The proposed method is based on the Faster RCNN baseline [20], pre-trained on the ImageNet. We optimize the network using the Stochastic Gradient Descent (SGD) algorithm with 0.9 momentum and 0.0005 weight decay, which is trained on 1 1080Ti GPU with the mini-batch involving 1 image per GPU. For Caltech dataset, we train the network for 40k iterations with the initial learning rate of 10^{-3} and decay it to 10^{-4} for another 20k iterations. For Citypersons dataset, we train the network for 20k iterations with the initial learning rate of 10^{-3} and decay it to 10^{-4} for another 10k iterations. All images are in the original 1x. scale during training and testing. Other parameters λ_1, λ_2,δ_0 and N_t are set to 1, 0.1,0.8 and 0.5 respectively.

B. Evaluation Metrics

• In experiments, we used the standard *average-log miss rate* (MR) on *False Positive Per Image* (FPPI) in $[10^{-2}, 10^0]$. This kind of metric is a bit similar to *Average Precision* (MAP) and refers more to the object not detected.

• On Caltech and Citypersons, we report results across two different occlusion degrees: Reasonable (**R**) and Heavy Occlusion (**HO**). The visibility ratio in **R** is lager than 65%, and the visibility ratio in **HO** ranges from 20% to 65%. In all subsets, the height of pedestrian over 50 pixels is taken for evaluation. It is worth noting that **HO** is designed to evaluate performance in case of severe occlusions.

TABLE I

COMPARISONS WITH THE STATE-OF-THE-ART METHODS WITH CHALLENGING SETTINGS ON CALTECH DATASET. RESULTS ARE THE MR^{-2} EVALUATION METRIC OF THE CORRESPONDING METHODS, IN WHICH LOWER IS BETTER. **BOLDFACE** INDICATES THE **BEST** PERFORMANCE. **GNMS** MEANS GREEDY NMS

Methods	Year	Backbone	R	HO
DeepParts [22]	2015	-	12.9	60.42
FasterRCNN+ATT [3]	2018	VGG16	8.11	45.18
MS-CNN [11]	2016	-	8.08	59.94
RPN+BF [23]	2016	-	7.28	74.36
SDS-RCNN [24]	2017	ResNet50	6.44	58.55
CA-GDFL [25]	2020	-	6.04	39.35
ALFNet [26]	2018	ResNet50	4.5	43.4
AR-Ped [27]	2019	ResNet50	4.36	48.80
Reploss [2]	2018	ResNet50	4.0	63.36
CSP [13]	2019	ResNet50	3.8	36.5
SA-DPM (GNMS)	-	ResNet50	4.02	32.66
SA-DPM (BNMS)	-	**ResNet50**	**3.45**	**30.18**

C. Ablation Study

In this section, we evaluate several issues that contribute to the overall performance under the standard MR^{-2} FPPI settings on Caltech dataset.

1) What Is the Role of Each Module: We conduct experiments of the individual modules compared with the proposed SA-DPM, respectively. The results are shown in Table I.

Performance of the Sequential Attention. We observer that (1) on one hand, compared with other anchor-based methods that are not specifically designed for occlusion detection, the proposed Sequential Attention achieves better performance on Caltech Heavy Occlusion subset (**HO**), e.g., AP-Ped (48.80% *vs.* 34.12%) and SDS-RCNN (58.55% *vs.* 34.12%); on the other hand, compared with some methods aimed at occluded pedestrian detection, such as FasterRCNN+ATT (45.18% *vs.* 34.12%) and DeepParts(60.42% *vs.* 34.12%), ours Sequential Attention still shows significant improvement. Besides, the proposed method without introducing any part-level annotations or any prior of part regions performs more flexible and accurate in the actual occlusion scenes; (2) despite anchor-free methods [13] have partially alleviated occlusion issues for their special point-based detection mechanism, their performance still falls behind us for the lack of oriented selection of body part features, e.g., CSP (36.5% *vs.* 34.12%). From the visual comparison in Figure 6, it can be seen that although the training process is under the weakly supervised manner, the proposed method can also perceive visible part features. However, we should not negate that this part-based detection also brings about more false positives (e.g., Figure 6, the red detection results without green GT boxes, in the first, fourth and fifth column of 2nd row). Therefore, it is necessary to introduce the Joint Learning, which not only learns where to focus but also learns how to aggregate various features to keep fewer FNs and FPs at the same time. Besides, it is

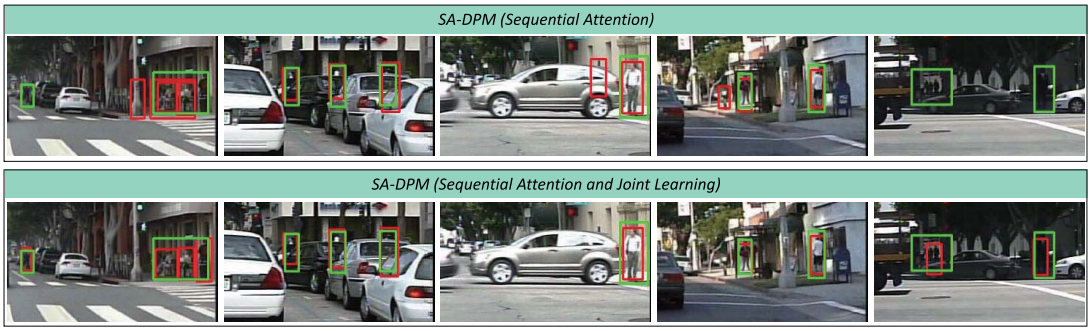


Fig. 6. Detection results on Caltech test set using the single Sequential Attention (upper row) and the proposed SA-DPM (lower row). The detection results are marked in red, while the ground truth are marked in green. The combination of two modules keeps lower FNs (the first and fifth column) and lower FPs (the first, third and fourth column).

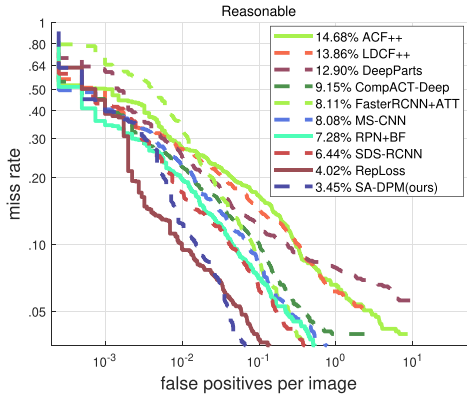


Fig. 7. Comparisons on the Caltech Reasonable subset, in which lower is better.

interesting to see that most informed regions are heads and arms, which to some extent, reflects the most recognized features of pedestrians are often in heads and arms.

Performance of the final representation. Table I shows that (1) either feature representation *alone* is useful for pedestrian detection. For instance, the Sequential Attention, which mainly utilizes visible part features, has achieved 34.12% on Caltech Heavy Occlusion subset. AR-Ped [27], which focus more on pedestrian full body features, still achieves well performance on general pedestrian detection (4.36% on Caltech Reasonable subset); (2) further performance gain is obtained by joining the two representations, yielding 3.94% (30.18% *vs.* 34.12%) on **HO** setting and 1.09% (3.45% *vs.* 4.54%) on **R** setting. It is noted that the joint learning manner not only brings improvement on general pedestrian detection, but also further enhances detection performance of occluded cases in MR^{-2} boost. This validates the complementary effect of jointly learning part and full-body features in our model. As is shown in Figure 6 (the first, fourth and sixth column), the error detection results (FNs and FPs) in the Sequential Attention could be corrected in the joint learning manner. Besides, we further visualize four parameters w^0, w^1, w^2 and w^3. As shown in Figure 8, if the pedestrian was not occluded, the final prediction features would be biased to full body. For occluded instances, the network tends to choose the most relevant part features. Besides, combined with the Sequential Attention, the Joint Learning improves all and obtains the best performance on both occluded and non-occluded instances. We compare our SA-DPM together with BNMS algorithm. As is shown in Table I, our SA-DPM with BNMS outperforms 0.57% with ResNet50 than the default setting with GNMS (Greedy NMS). Furthermore, our model still achieves 2.48% on **HO** set, which demonstrates that the BNMS performs better post-processing in the challenging scenes and keeps more accurate detection results.

TABLE II

THE ABLATION STUDY OF TWO KINDS OF PART LOCATOR UNDER THE MR^{-2} METRIC. CAT AND SAT REPRESENT CHANNEL ATTENTION AND SPATIAL ATTENTION, RESPECTIVELY

	CAT	SAT	R	HO
baseline			7.21	44.34
SA-DPM	✓		5.91	38.66
SA-DPM		✓	4.55	33.89
SA-DPM	✓	✓	**3.45**	**30.18**

2) Spatial Attention vs. Channel Attention: Spatial and channel attention are two conceptually different mechanisms. (1) spatial attention deals with all channels in the same way on a 2D plane. That is, it has the same weight on all channels, while for an individual plane, the weights are different. spatial attention focuses more on pedestrian structure information, which means pedestrians always act out up-straight posture, and the body parts are distributed from top to bottom. (2) On contrast, the weights of channel attention are the same on the plane but different across channels. The separate results of spatial and channel attention are shown in Table II. We can discover that (1) they both bring gain of pedestrian detection performance. (2) the combination of spatial and channel attention gives further accuracy boost, which suggests the complementary information between the two kinds introduced by our model. This shows that our method is effective in identifying and exploiting the complementary information of pedestrian features.

3) The Number of Steps: **The number of steps** We also analyze the impact of the hyper-parameters, namely the number of steps (K), on two datasets. Figure 10 presents the results across the number of steps in the Sequential Attention, where performances increase until the certain number ($K = 3$ for

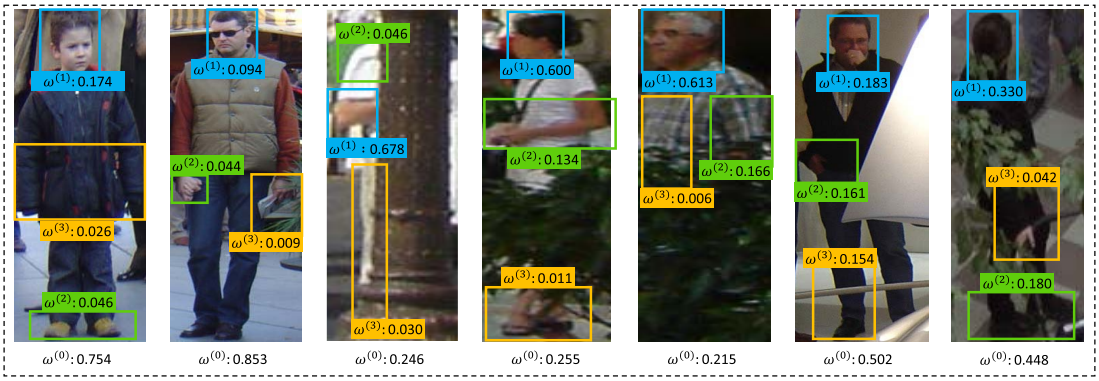


Fig. 8. The part detection boxes of the proposed sequential attention module and the corresponding weights w^0, w^1, w^2, w^3 predicted in the Joint Learning.

Fig. 9. Detection results of the proposed method. The most relevant part detection results are marked in dotted red, while the overall detection results are marked in green. Note that from the results above, the proposed method can adaptively select part or full body to predict pedestrian instances.

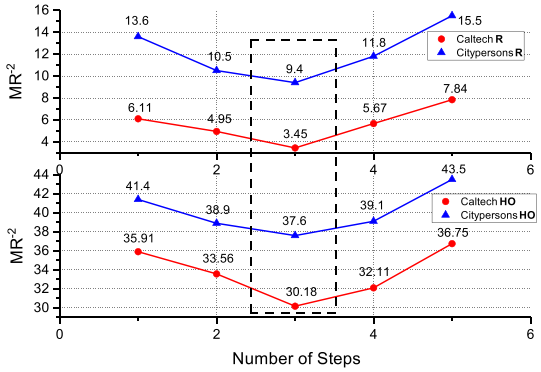


Fig. 10. Ablation study of different number of steps under the MR^{-2} metric.

both Caltech and Citypersons datasets) and decrease afterwards. This is because under weakly-supervised manner, larger K makes models difficult to capture accurate part regions, and thus the incorrect part features could prevent us from training an accurate pedestrian detector. As shown in Figure 8, when the number of steps is set to the constant 3, each step can extract accurate part region. Besides, under the constraint of Frobenius norm, the semantic meaning of part features extracted from various steps is different, in which the first step usually focus on head regions, while the second and third step pay more attention to hands and lower-body regions. This is consistent with our experience that in different scenes, the head region is the least likely to be occluded and is the most distinct feature to identify the pedestrian instance.

D. Caltech Dataset

The Caltech Dataset [9] consists of approximately 250,000 frames with a total of 350,000 bounding boxes and 2300 unique pedestrians were annotated. We evaluate on the standard 4024 images using the **R** and **H** set. The proposed method is extensively compared with the state-of-the-arts. As is shown in Table I, SA-DPM achieves 0.45% on the **R** set of CSP [13], 1.05% of ALFNet [26] and 2.99% of SDS-RCNN [24]. it presents the superiority on detection pedestrians without occlusion. Moreover, Table I shows that SA-DPM also performs very well for heavy occlusion instances, which outperforms 6.32% on the **HO** set of CSP, 13.22% of ALFNet and 15.00% of FasterRCNN+ATT [3].

TABLE III

COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON CITYPERSONS VALIDATION DATASET. RESULTS ARE THE MR^{-2} EVALUATION METRIC OF THE CORRESPONDING METHODS, IN WHICH LOWER IS BETTER. **BOLDFACE** INDICATES THE **BEST** PERFORMANCE

	Year	**Backbone**	**R**	**Time**
FRCNN [3]	2018	VGG16	15.4	-
OR-CNN [4]	2018	VGG16	12.8	-
Bi-box [16]	2018	VGG16	11.2	-
PBM [7]	2020	VGG16	11.1	-
EMD [29]	2020	VGG16	10.7	-
MGAN [8]	2019	VGG16	10.5	-
SA-DPM (ours)	-	VGG16	10.9	0.24s/img
TLL [30]	2018	ResNet50	15.5	-
TLL+MRF [30]	2018	ResNet50	14.4	-
RepLoss [2]	2018	ResNet50	13.2	-
ALFNet [26]	2018	ResNet50	12.0	0.27s/img
CSP [13]	2019	ResNet50	11.0	0.33s/img
STDA+ [15]	2021	ResNet50	10.0	-
SA-DPM (ours)	-	**ResNet50**	**9.7**	**0.16s/img**

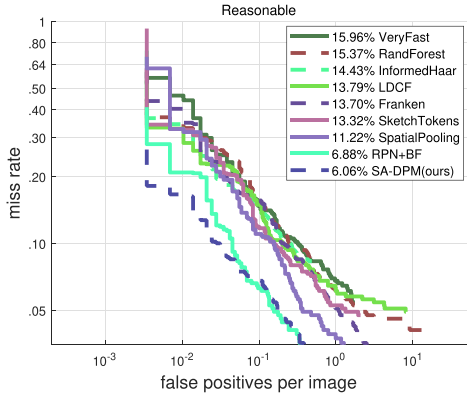


Fig. 11. Comparison results on the INRIA dataset, in which lower is better.

E. Citypersons Dataset

Citypersons [10] is a diverse dataset built upon the Cityscapes data, which includes 5000 images (2975 for training, 500 for validation, and 1525 for testing). In a total of 5 000 images, it has ∼35k person and ∼13k ignore region annotations. And it notices the density of persons are consistent across train/validation/test subsets. From Table III, it can be observed that SA-DPM beats the competitors and performs fairly well on occlusion cases. On the **R** set, SA-DPM with BNMS achieves the best performance, with a competitive inference speed 0.16s/img.

F. INRIA Dataset

INRIA provides original pictures and corresponding annotation files. There were 614 positive samples (2416 pedestrians) and 1218 negative samples in the training set, 288 positive samples (1126 pedestrians) and 453 negative samples in the test set. Most of the human body in the picture is in standing posture and its height is more than 100 pixels. We use the 614 positive images during training to fine-tune our model pre-trained on Caltech. Figure 11 shows the proposed SA-DPM leads a new state-of-the-art result of 6.06% MR^{-2} on INRIA dataset, which verifies the effectiveness of the proposed method.

V. CONCLUSION

In this work, we constructed a Sequential Attention based Distinct Part Model (SA-DPM) for occluded pedestrian detection. In order to enhance the whole feature representation, we employed distinct part perception (Sequential Attention) to extract part-level features and balanced re-weighting (Joint Learning) to seize the relationship between part and full-body features adaptively. In addition, to alleviate the dilemma of Greedy NMS in setting the IoU threshold, we designed the BNMS module using the above detected part and full-body boxes as inputs. Our method is trained in an end-to-end fashion and achieves the state-of-the-art accuracy on several widely used datasets Caltech and Citypersons. In addition to its well performance on the Reasonable subset, our method has better performance for occluded pedestrian detection. It is of more importance for the practical application of pedestrian detection. Although the performance is better, there are also certain limitations, mainly focusing on the complex modules and excessive parameters. We would like to explore one more lightweight pedestrian detector in future.

REFERENCES

[1] J. Cao, Y. Pang, J. Han, B. Gao, and X. Li, "Taking a look at small-scale pedestrians and occluded pedestrians," *IEEE Trans. Image Process.*, vol. 29, pp. 3143–3152, 2020.

[2] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7774–7783.

[3] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6995–7003.

[4] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.

[5] H. Cheng, C. Zhang, W. Song, Y. Li, and Y. P. Zhong, "Pedestrian detection with multi-scale context-embedded feature learning," in *Proc. 4th IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2017, pp. 346–351.

[6] Z. Liu, C. Zhang, Y. Luo, K. Chen, Q. Zhou, and Y. Lai, "Improving small-scale pedestrian detection using informed context," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 1–4.

[7] X. Huang, Z. Ge, Z. Jie, and O. Yoshie, "NMS by representative region: Towards crowded pedestrian detection by proposal pairing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10750–10759.

[8] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-guided attention network for occluded pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4967–4975.

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 304–311.

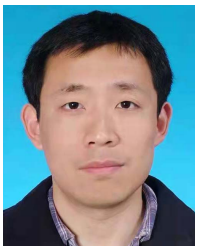
[10] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3213–3221.

[11] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 354–370.

[12] X. Du, M. El-Khamy, J. Lee, and L. Davis, "Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 953–961.
[13] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5187–5196.
[14] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[15] Z. Chen, W. Ouyang, T. Liu, and D. Tao, "A shape transformation-based dataset augmentation framework for pedestrian detection," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1121–1138, Jan. 2021, doi: 10.1007/s11263-020-01412-0.
[16] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 135–151.
[17] Y. Zhao, Z. Yuan, and B. Chen, "Accurate pedestrian detection by human pose regression," *IEEE Trans. Image Process.*, vol. 29, pp. 1591–1605, 2020.
[18] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008.
[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[21] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Relational learning for joint head and human detection," 2019, *arxiv 1909.107674*.
[22] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1904–1912.
[23] L. Zhang, L. Liang, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 443–457.
[24] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection & segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4950–4959.
[25] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for robust pedestrian detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3820–3834, 2020.
[26] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 618–634.
[27] G. Brazil and X. Liu, "Pedestrian detection with autoregressive network phases," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7231–7240.
[28] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[29] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in crowded scenes: One proposal, multiple predictions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12214–12223.
[30] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 554–569.

Yan Luo received the B.S. degree in information engineering from Southeast University, China. She is currently pursuing the Ph.D. degree in information and communication engineering with Shanghai Jiao Tong University, China. Her research interests are focused on pedestrian detection, pattern recognition, machine learning, computer vision, and intelligent transportation systems.

Chongyang Zhang (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2008. He is currently a Professor with the Department of Electronic Engineering, Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University. His research interests are in the areas of machine learning and computer vision, especially on object detection, crowd counting, action recognition, and event detection. He has published over 50 international journals or conference papers on these topics.



Weiyao Lin (Senior Member, IEEE) received the B.S. and M.E. degrees in electrical engineering from Shanghai Jiao Tong University in 2003 and 2005, respectively, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, in 2010. From 2006 to 2010, he worked as a Research Intern at Motorola Inc., RealNetworks, and Thomson Technology. In 2010, he joined the Department of Electronic Engineering, Shanghai Jiao Tong University, where he is currently a Professor.

Xiaokang Yang (Fellow, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2000. He is currently a Distinguished Professor with the School of Electronic Information and Electrical Engineering and the Deputy Director of the Artificial Intelligence Institute, Shanghai Jiao Tong University. He has published over 200 refereed papers, and has filed 60 patents.

Jun Sun (Member, IEEE) received the B.S. degree in electrical engineering from the University of Electronic Sciences and technology of China, Chengdu, China, in 1989, and the Ph.D. degree in electrical engineering from Shanghai Jiao Tong University in 1995. He is currently a Professor and a Ph.D. Advisor at Shanghai Jiao Tong University. He has published over 50 technical articles in the areas of digital television and multimedia communications. His research interests include digital television, image communication, and video encoding.