DDAD: Detachable Crowd Density Estimation Assisted Pedestrian Detection

Wenxiao Tang, *Student Member, IEEE*, Kun Liu, M. Saad Shakeel, Hao Wang, and Wenxiong Kang, *Member, IEEE*

***Abstract*— Detecting pedestrians is a challenging computer vision task, especially in the intelligent transportation system. Mainstream pedestrian detection methods purely utilize information of bounding boxes, which overlooks the role of other valuable attributes (e.g., head, head-shoulders, and keypoints) of pedestrians and leads to sub-optimal solutions. Some works leveraged these valuable attributes with a minor performance improvement at the expense of increased computational complexity during the inference phase. To alleviate this dilemma, we propose a simple yet effective method, namely Detachable crowd Density estimation Assisted pedestrian Detection (DDAD), which leverages the crowd density attributes to assist pedestrian detection in the real-world scenes (e.g., crowded scenes and small-scale pedestrian scenes). The advantage of the crowd density estimation is that it allows the network to focus more on the human head and the small-scale pedestrians, which improves the features representation of pedestrians heavily occluded or far from cameras. Our DDAD works on a principle of multi-task learning and can be seamlessly applied to both one-stage and two-stage pedestrian detectors by equipping them with an extra detachable branch of crowd density estimation. The equipped crowd density estimation branch is trained with the annotations derived from the existing pedestrian bounding box annotations, occurring no extra annotation cost. Moreover, it can be removed during the inference phase without sacrificing the inference speed. Extensive experiments conducted on two challenging datasets, i.e., CrowdHuman and CityPersons, demonstrate that our proposed DDAD achieves a significant improvement upon the state-of-the-art methods. Code is available at https://github.com/SCUT-BIP-Lab/DDAD.**

***Index Terms*— Pedestrian detection, multi-task learning, crowd density estimation.**

Manuscript received 20 February 2022; revised 23 August 2022; accepted 1 November 2022. Date of publication 24 November 2022; date of current version 8 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61976095 and in part by the Natural Science Foundation of Guangdong Province, China, under Grant 2022A1515010114. The Associate Editor for this article was S. Hamdar. *(Corresponding authors: Hao Wang; Wenxiong Kang.)*

Wenxiao Tang, Kun Liu, and M. Saad Shakeel are with the School of Automation Science and Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: autang_wx@mail.scut.edu.cn).

Hao Wang is with the School of Automation Science and Engineering, South China University of Technology, Guangzhou 510641, China, and also with Guangdong Airport Baiyun Information Technology Company Ltd., Post-Doctoral Innovation Practice Base, Guangzhou 510470, China (e-mail: meeric_wan@mail.suct.edu.cn).

Wenxiong Kang is with the School of Automation Science and Engineering, South China University of Technology, Guangzhou 510641, China, also with the Pazhou Laboratory, Guangzhou 510335, China, and also with the Guangdong Enterprise Key Laboratory of Intelligent Finance, Guangzhou 510705, China (e-mail: auwxkang@scut.edu.cn).

Digital Object Identifier 10.1109/TITS.2022.3222692

I. Introduction

THE pedestrian detection is a challenging computer vision task with numerous real-world applications, such as autonomous driving [1], intelligent surveillance [2] and robotics [3], which makes it an important component of the intelligent transportation system. Recently, Deep Convolution Neural Networks (DCNNs) [4], [5], [6], [7] have demonstrated remarkable performance gains compared to traditional approaches [8], [9], [10] in pedestrian detection. However, these methods simply utilize the pedestrian frame, discarding the other valuable pedestrian attributes (e.g., head, head-shoulders, segmentation, and keypoints), which makes their performance unsatisfactory, especially when detecting small-scale pedestrians far from the camera in crowded scenarios.

To fully exploit the valuable pedestrian attributes, Hou and Pang [11] leveraged the number of persons in a given scene as the prior information for the detection task, which ignores the valuable local information. Some recent studies [12], [13] developed a collaborative network for head and pedestrian detection. Similarly, other approaches [14], [15] proposed to detect visible part and full body of the pedestrian simultaneously, which are then combined to improve the pedestrian detection performance. Besides, He et al. [16] and Jiao et al. [17] proposed to utilize segmentation and pose estimation model to assist pedestrian detection tasks, respectively. Despite the fact that the above-mentioned methods [12], [13], [14], [15], [16], [17] have made significant efforts in crowded pedestrian detection, they still suffer from two critical drawbacks: 1) They require extra manually labeled annotations for the complementary task learning, significantly increasing the annotation cost; 2) They bring a large computation overhead in the inference phase, which impedes their deployment in real-time application scenarios. Therefore, there is still room for research in exploiting the valuable pedestrian attributes to improve the pedestrian detection performance without additional manual annotation cost and computational burden in the training and inference phases, respectively.

In this work, we propose a simple but highly-efficient pedestrian detection approach named Detachable crowd Density estimation Assisted pedestrian Detection (DDAD), without any extra annotation cost in the training phase and computational complexity in the inference phase. Our DDAD is formulated as a multi-task learning manner by simply equipping the pedestrian detector with a detachable branch for crowd density

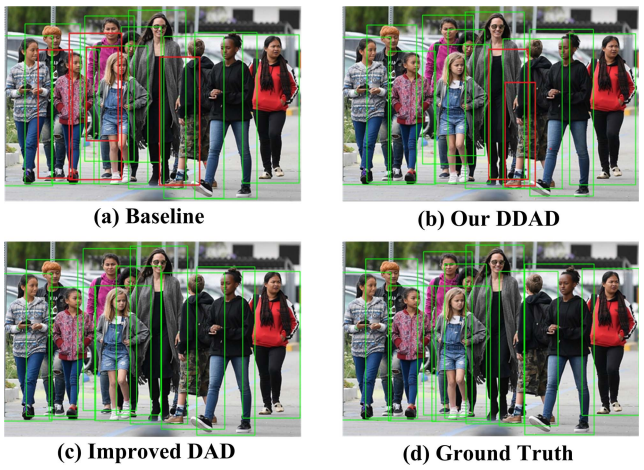


Fig. 1. Pedestrian detection in crowded scenes. (a) Detection results are predicted by the FPN baseline, where the solid red boxes indicate error detection. (b) Detection results of our DDAD based on FPN. (c) Detection results of our Improved DAD, where all the pedestrians are correctly detected. (d) The ground truth of the image.

estimation, which can be seamlessly applied to both one-stage and two-stage ones. The annotations required for the crowd density estimation branch are generated automatically from the existing pedestrian bounding boxes, which incurs no extra cost for manual labeling. Specifically, we use the midpoint of the upper boundary of the pedestrian bounding box to approximate the annotation for the crowd density estimation task. Besides, the crowd density estimation branch is detachable during the inference phase, which prevents it from sacrificing the inference speed of the vanilla pedestrian detector.

Furthermore, to fully explore the collaboration between these two pedestrian-related tasks, we gradually integrate features from the deep layers of crowd density estimation branch into the pedestrian detection branch, which results in the Improved crowd Density estimation Assisted pedestrian Detection (IDAD). Our experiments show that the IDAD gains further performance improvement compared to DDAD, as some visualization results are illustrated in Fig. 1. We also demonstrate that the performance improvement by IDAD is mainly due to the cooperative nature of these two pedestrian-related tasks. The experiment results show that our designed multi-task learning framework is rational and effective.

To sum up, we make the following contributions:

- We propose a multi-task learning framework, named DDAD, by equipping vanilla pedestrian detectors with a detachable crowd density estimation branch, which can be seamlessly applied to both one-stage and two-stage pedestrian detectors.
- A novel method for generating the approximate crowd density annotation from existing pedestrian bounding boxes annotation is proposed, with which our DDAD achieves a mean absolute error (MAE) of 4.3 on the CrowdHuman [18] dataset.
- The proposed DDAD can significantly improve the performance of existing mainstream pedestrian detection methods without extra annotation and computation burden in the training and inference phases, respectively.
- We demonstrate the cooperative nature of these two pedestrian-related tasks with a practical feature fusion on their deep layer, named IDAD, which further improves pedestrian detection performance with a small overhead.

The rest of this paper is organized as follows. Section II reviews some related works on crowd density estimation and pedestrian detection. Section III presents the details of our Detachable crowd Density estimation Assisted pedestrian Detection (DDAD) framework, followed by the Improved crowd Density estimation Assisted pedestrian Detection (IDAD). Section IV presents the experimental results with comprehensive comparative analysis, ablation studies, and a thorough discussion. Section V concludes the paper with the discussion on potential future works.

II. Related Work

A. Crowd Density Estimation

Following the idea proposed in [19], crowd density estimation methods can be divided into three categories: detection-based methods, regression-based methods, and density-based methods. The majority of the early approaches [20], [21], [22] detect pedestrians, heads, or upper bodies in images and count the number of them. However, these methods can not perform well in crowded scenes where pedestrians are obscured by others, making it difficult to correctly detect an occluded head. In addition, they required manually annotate bounding boxes. Although some methods [23], [24], [25] directly regressed the number of pedestrians from the feature vector, they completely ignored the spatial information and do not make proper use of the head-point annotation maps. These problems existed until the introduction of the crowd density map estimation method [26]. Most existing methods [27], [28], [29] utilized crowd density map estimation and were therefore more robust than the first two categories of methods mentioned above. Zhang et al. [26] proposed a domain-adaptive crowd density estimation algorithm that trained a crowd density estimator with ground truth annotations in source scenes and then approximately annotated some local patches from the target domain with similar properties to those of the source domain. However, these methods required separate labeling of the crowd density estimation annotations to achieve satisfactory performance. In contrast, we propose a novel alternative approach to generating annotations for the crowd density estimation task, which approximately takes the midpoint of the upper boundary of the pedestrian bounding box as the head point annotation.

B. Pedestrian Detection

Traditional pedestrian detectors [8], [9], [10] extracted color, texture, and edge information were always unsatisfactory in real-world applications. With the success and popularity of Faster R-CNN [30] for generic object detection, deep learning-based pedestrian detection methods have achieved better performance. Wang et al. [31] designed a joint semantic segmentation and pedestrian detection algorithm to distinguish foreground from background for better pedestrian detection performance. Kishore et al. [32] and Zhao et al. [33] incorporated pedestrian detection and human pose estimation into

a cascade structure. Some recent methods [12], [13], [14] have made progress in crowded scenes by utilizing extra annotations, such as head annotations or visible annotations as an additional supervision signal. Specifically, Chi et al. [12] and Zhang et al. [13] utilized head annotations to assist full-body detection in crowded scenes. They detected the full body assisted by head detection, which introduced complex computations in the inference phase. It is not coincidental that PedHunter [34], JointDet [12], MGAN [35], and PDOE [14] utilized visible body information or head information to improve the performance of pedestrian detection. Although these methods exploited useful pedestrian attribute beyond the pedestrian frames and improve the performance of pedestrian detection models partially, they required extra annotations such as segmentation labels, key-points, head, or visible bounding boxes, which leads to extremely high annotation costs. What's more, most of them slow down the inference speed of pedestrian detection. In this paper, our proposed DDAD fully exploits an additional attribute of pedestrians, crowd density, to achieve better performance. In contrast, our DDAD requires only full-body annotation without any extra annotations and, will not sacrificing the inference speed. A similar work to ours is that Rodriguez et al. [36] proposed to utilize global crowd density information to refine the coarse head detection, which can only be used for head detection, but not tailored for pedestrian detection. In their method, the annotations for crowd density estimation were obtained from the head bounding box. Moreover, their method requires crowd density information to predict head bounding boxes in the inference phase, which slows down their inference speed.

C. Feature Fusion in Computer Vision

The purpose of feature fusion is to capture useful contextual and semantic information from different layers of the deep network. Ren et al. [37] built a recurrent rolling convolutional architecture that combines features from different CNN layers to generate high-level features. Shang et al. [38] introduced a complementary subnetwork to generate the high-resolution feature map for small-scale object detection. Zhang et al. [39] concatenated the RoI features from different layers along with the global context. Wang et al. [40] proposed a local competition mechanism (maxout) for adaptive context fusion. Cao et al. [41] embedded the large-kernel convolution into feature pyramid structure to exploit contextual information. Wu et al. [5] proposed to adapt features from the current and nearby frames to achieve robust pedestrian detection. However, these methods perform feature fusion for only one particular task. On the contrary, our IDAD fuses feature maps from different tasks to generate high-level features in several consecutive layers, which is more challenging and requires a higher degree of collaborative nature between the two tasks.

III. Proposed Approach

In this work, we mainly focus on efficiently utilizing extra valuable pedestrian attributes to improve pedestrian detection performance without increasing the annotation cost and inference time. Our insight of DDAD comes from the observation of people annotating pedestrian in the image. When we annotate the pedestrian, we quickly browse the image for the global information that we consider to be valuable attributes of the pedestrian. Among these attributes, e.g., head location, pose, visible area, and crowd density, the most quickly accessible is crowd density. A coarse hierarchy of crowd density in images, such as no occlusion, slight occlusion and severe occlusion, is also performed in our minds. Only after that, we will start annotating the pedestrian. In other words, each pedestrian in the image is localized after the annotator yields rough knowledge about its crowd density. Therefore, if the pedestrian detector has a coarse knowledge of the crowd density in the image, it can focus more on the upper boundary region of the pedestrians, which naturally detects each pedestrian with a higher accuracy. To this end, we propose a simple yet effective method, namely Detachable crowd Density estimation Assisted pedestrian Detection (DDAD), which leverages the crowd density attributes to assist pedestrian detection. In this section, we elaborate our proposed method from five aspects: (1) The structure of DDAD, (2) The crowd density estimation network, (3) Generation of crowd density estimation annotations from the existing pedestrian bounding boxes, (4) Feature fusion strategy for these two pedestrian-related tasks, (5) Training and inference implementation.

A. The Structure of DDAD

Our DDAD works on a principle of multi-task learning and can be seamlessly applied to both one-stage and two-stage pedestrian detectors by equipping them with a detachable crowd density estimation branch. The overall architecture of our proposed DDAD is shown in Fig. 2, which comprises a backbone network, a pedestrian detection branch, and a crowd density estimation branch. The pedestrian detection branch can adopt any of the current mainstream pedestrian detectors. In this work, we choose Feature Pyramid Network (FPN) [42], CrowdDet [43], CSP [44] as our baseline models for pedestrian detection. For more detailed information, we suggest the readers refer to the original papers [42], [43], [44]. It should be noted that our proposed DDAD primarily improves the pedestrian detection performance in complicated and severely occluded outdoor scenes. The change in the distance between the pedestrian and the camera can cause a significant scale change. Therefore, we choose FPN [42] as one of our baselines due to its ability to extract multi-scale information. So it can make it easier to detect pedestrians at different distances in outdoor scenes. As for the other two baseline pedestrian detection approaches, CSP [44] and CrowdDet [43], are recognized as highly effective one-stage and two-stage pedestrian detectors, respectively. Furthermore, to make our proposed DDAD feasible for real-world applications, we also validate the effectiveness of our proposed DDAD with YOLO v5s [45].

In the literature of crowd density estimation, many state-of-the-art frameworks [46], [47] follow the design of U-Net [48], whose core principle is to gradually fuse high-level features with low-level ones to generate high-resolution and informative crowd density maps. Therefore, in this work, we simply utilize the U-Net [48] to estimate the crowd density, which will be described in the following part. The branch of

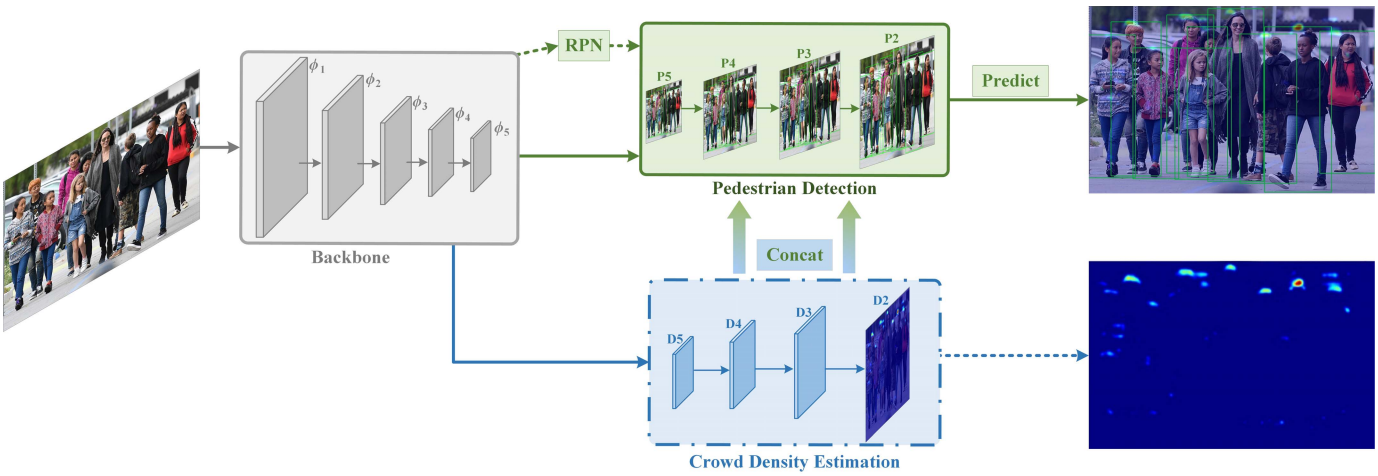


Fig. 2. The architecture of our proposed DDAD, which comprises a backbone network, a pedestrian detection branch, and a crowd density estimation branch. Both branches share the same backbone network. The branch of pedestrian detection can be both one-stage and two-stage pedestrian detectors. The main difference between them is the presence or absence of the Region Proposal Network (RPN) structure. The crowd density estimation branch will be removed during the inference of our proposed DDAD method, while it will be retained in our Improved DAD.

pedestrian detection and crowd density estimation share the same backbone, expecting to capture pedestrian information as well as crowd density information. We utilize ResNet-50 [49] as our backbone network for fair comparison, which is consistent with many previous pedestrian detectors [42], [43], [44]. It should be noted that the ResNet architecture can not only facilitate the fast and accurate convergence of the convolution neural networks but also avoid overfitting when constructing a deeper network.

B. The Crowd Density Estimation Network

Without loss of generality, we take the vanilla pedestrian detector FPN [42] as an example to introduce the crowd density estimation branch in our DDAD. A simple U-Net network is used to estimate the crowd density map. We denote the output of our backbone network from 2^{nd}, 3^{rd}, 4^{th} and 5^{th} stage as ϕ_2, ϕ_3, ϕ_4 and ϕ_5 respectively, where the feature maps from the shallow layers contain more precise localization information, while the deep layers extract rich semantic information with large receptive fields. Therefore, we choose feature maps from multiple stages as an input for our crowd density estimation branch. The output of the density estimation branch is defined as follows:

$$\phi_{\text{density}} = f_2\left(\phi_2 + f_3\left(\phi_3 + f_4\left(\phi_4 + f_5\left(\phi_5\right)\right)\right)\right) \quad (1)$$

where ϕ_i represents the feature map output from the i^{th} stage, and f means dimension-reduction and up-sampling operation based on bilinear interpolation. Specifically, taking the image as an input, the backbone network may generate several feature maps with different resolutions and dimensions. We first reduce its dimension of the 5^{th} stage from 2048 to 256 using 1×1 convolutional layer to decrease computational complexity, followed by a bilinear interpolation-based up-sampling operation and pixel-wise addition to the shallower one. This operation will be continued until the resolution of feature maps is $1/4$ of the original input image.

C. Ground Truth for Crowd Density Estimation

As explained in the above subsection, we follow a density-based estimation method and design a simple U-Net to train the crowd density estimation network. However, the common pedestrian detection datasets exclude the annotations for crowd density estimation. Although the CrowdHuman [18] dataset contains pedestrian head bounding box annotations, it can not directly serve as an annotation for crowd density estimation task. Additionally, the CityPersons [50] dataset does not contain the head bounding box annotations, which makes it more challenging to train a crowd density estimation network. Therefore, it is crucial to reasonably convert the exist annotations from the pedestrian detection dataset to that for the crowd density estimation task.

In this section, we elaborate on our proposed solution to overcome such limitations. Firstly, we utilize the midpoint of the head bounding box as an annotation for the crowd density estimation branch, which is an easy way to alleviate the lacking of crowd density estimation annotation. However, the annotation of the midpoint of the head bounding boxes and that of the crowd density estimation is somewhat similar but not identical. We conduct an experiment to generate the annotations required for density estimation using the midpoint of the head annotation box and achieve satisfactory pedestrian detection performance, which will be shown in the Experiments section. However, what hinders our further research is that some other pedestrian detection datasets (*e.g.*, CityPersons [50]) excludes the head bounding box annotation. In practical applications, it is quite expensive to obtain head boxes annotations for crowd density estimation. The question arises that can we replace the point annotations needed for the crowd density estimator with some other existing annotations in the pedestrian detection datasets? In this regard, we resort to the full-body annotation required for pedestrian detection and utilize the midpoint of the upper boundary of the full-body box as an annotation for the crowd density estimation network. To the best of our knowledge, our method is the first one

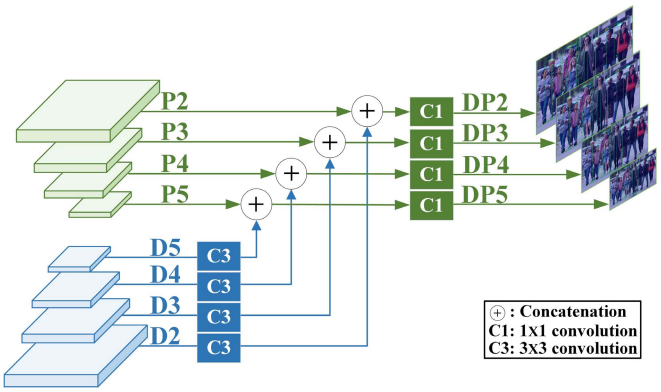


Fig. 3. The structure of our IDAD. We further concatenate the different layers features from the crowd density estimation branch (D) and the pedestrian detection branch (P). The fused features are dimensionality reduced with Conv and then used for pedestrian detection.

that trains crowd density estimation network with only pedestrian full-body bounding boxes annotations, without using any crowd density estimation annotations.

Generally thinking, we convert the midpoint of the upper boundary of the full-body bounding boxes into a crowd density map in a common way. Specifically, if there is a point at pixel x_i, we represent it as a delta function $\delta(x - x_i)$. Hence an image with N labeled points can be represented as:

$$H(x) = \sum_{i=1}^{N} \delta(x - x_i) \tag{2}$$

To convert Eq.2 into a continuous density function, we convolve it with a Gaussian kernel $G_\sigma(x)$. Hence, the ground truth of crowd density is defined as follows:

$$F(x) = H(x) * G_\sigma(x) \tag{3}$$

D. Deep Feature Fusion Strategy for Collaborative Tasks

Our proposed DDAD is formulated as a multi-task learning framework, where the pedestrian detection task and crowd density estimation task are trained together with the same backbone network. In the inference phase, we detach the crowd density estimation branch to keep the same inference speed as the vanilla pedestrian detector but significantly improve the pedestrian detection performance. In the case of correlation between different tasks in a multi-task learning framework, the overall performance of the system can be improved by leveraging from each other's valuable attributes or information. Taking our DDAD as an example, we enhance the pedestrian detection performance by adding an additional branch of crowd density estimation during training. We believe that these two tasks (i.e., pedestrian detection and crowd density estimation) are somewhat correlated.

To further explore the collaboration of these two pedestrian-related tasks, we proposed an Improved crowd Density estimation Assisted pedestrian Detection (IDAD) framework that fuses the high-level features of these two branches in deep layers, which are then used for pedestrian detection task. Fig. 3 illustrates the core structure of our IDAD. Specifically, we define the output feature maps from the i^{th} stage of the crowd density estimator and the pedestrian detector as D_i and P_i, respectively. Thus, the features used for pedestrian detection after the fusion strategy can be represented as:

$$DP_i = f_i^{\omega}\left(f_i^{\theta}(D_i) \bigoplus P_i\right) \tag{4}$$

where f_i^{θ} is a 3×3 convolution used to convert the crowd density estimation features for fusion with the pedestrian detection features, $\bigoplus$ represents the concatenation operation of different features, and f_i^{ω} is a 1×1 convolution operator. Similar to the vanilla model of pedestrian detector, the fused features are used as an input to the head of the pedestrian detection.

E. Training and Inference

1) Training: The whole network is optimized by the weighted loss L, which can be written as:

$$L = \lambda_1 L_{cls} + \lambda_2 L_{box} + \lambda_3 L_{num} + \lambda_4 L_{density} \tag{5}$$

where λ_i is the coefficient used to balance the gradient magnitudes of all the corresponding loss functions. The classification loss L_{cls} and the regression loss L_{box} are identical to those defined in [51]. The crowd density loss $L_{density}$ and the total number loss L_{num} correspond to the MSE loss. The crowd density loss $L_{density}$ measures the error between the predicted output ϕ_{density} in Eq.(1) and the ground truth heatmap $F(x)$ in Eq.(3), and the total number loss L_{num} calculates the error between the predict number of pedestrian and that of the ground truth.

2) Inference: In the inference phase, our proposed DDAD removes the crowd density estimation branch and retain the branch for pedestrian detection. Therefore, it keeps the consistent inference speed compared to baseline pedestrian detector without any extra overhead. Further, we delved into the collaborative property of these two pedestrian-related tasks and proposed IDAD. Our proposed IDAD that fuses the deep features of the crowd density estimation branch and the pedestrian detection branch at multiple layers with negligible extra computation in the inference phase.

IV. Experiments

In this section, we compare our proposed method with four baseline pedestrian detection frameworks (i.e., FPN [42], CrowdDet [43], CSP [44], and YOLO v5s [45]) to demonstrate the effectiveness of our proposed DDAD on two challenging benchmark datasets: the CityPersons [50] dataset and the CrowdHuman [18] dataset. In general, pedestrian detection algorithms tend to recall more instances in crowded scenes for better performance, which leads to an increased risk of false-positive predictions. In pursuit of improving both the opposite aspects, our proposed DDAD outperforms the baseline pedestrian detectors and sets a new state-of-the-art on both datasets. For the convenience of expression and comprehension, we utilize the notations "+" and "++" to denote the DDAD and the IDAD, respectively. It should be noted that both branches, including the pedestrian detection branch and

TABLE I
STATISTICS OF THE TWO MOST DOMINANT PEDESTRIAN DETECTION DATASETS IN TERMS OF NUMBER OF IMAGES, CROWD DENSITY, ETC. WE ONLY SHOW THE STATISTICS OF TRAINING SUBSET

	# images	# persons	# ignore regions	# person/image	# overlaps/images
CityPersons [50]	2,975	19,238	6,768	6.47	0.32
CrowdHuman [18]	15,000	339,565	99,227	22.64	2.40

crowd density estimation branch, are trained simultaneously in all of our experiments. Furthermore, for a fair comparison, the data augmentation strategy remains consistent with the baseline pedestrian detectors.

A. Datasets and Evaluation Metrics

1) DataSets: An ideal pedestrian detector should be robust to a variety of test-case scenarios, i.e., it should not only be effective in detecting pedestrians in crowded scenarios but also in single/uncrowded ones. To demonstrate the effectiveness of our proposed DDAD, we use two of the most popular and challenging pedestrian detection benchmark datasets, Citypersons [50] and CrowdHuman [18], both containing many challenging scenarios. The complete statistics of these two training subsets are shown in TABLE I. The CityPersons [50] dataset contains 2975 training images, 500 validation images, and 1575 test images, captured in three different seasons under various weather conditions. The crowdHuman [18] is a recently released dataset to evaluate pedestrian detection performance in crowded scenes. It contains 15,000 training images, 4370 validation images, and 5,000 test images captured from different cameras at different angles. Following [43], we train our pedestrian detector on the provided training set and test it on the validation set for a fair comparison.

It should be noted that the core objective of our proposed DDAD is to improve the performance of pedestrian detection in crowded scenes. Therefore, we conduct most of the experiments on the CrowdHuman [18] dataset, which is the most challenging dataset in the pedestrian detection literature, with the highest average number of pedestrians. Meanwhile, we also conducted experiments on the CityPersons [50] dataset, which contains fewer high density pedestrian areas than the CrowdHuman [18] dataset. This allows us to evaluate the effectiveness of our DDAD in different scenarios. In the field of pedestrian detection, the CityPersons [50] dataset ranks the degree of occlusion which can be computed as:

$$OCC_{rate} = 1 - \frac{\text{area}\,(\text{Bbox}_{\text{visible}})}{\text{area}(\text{Bbox})} \tag{6}$$

where the $\text{Bbox}_{\text{visible}}$ and Bbox represent the visible and full pedestrian body, respectively. The occlusion rate OCC_{rate} is higher when the visible area of pedestrians is small, which can also be interpreted as a large obscured area. In particular, the OCC_{rate} of less than 0.35 is considered the subset with reasonable occlusion, while it between 0.35 and 0.75 is considered the heavy occluded subset, and when it greater than 0.75 is considered the very high occluded subset. For more detailed information, please refer to the vanilla paper [50].

2) Evaluation Metrics: In our experiments, we mainly use the following three evaluation metrics for performance evaluation:

- *Average Precision* (AP) is the most popular evaluation metric for pedestrian detection task, which is computed by taking the average of the highest precision under different recalls. It reflects the precision and recall ratio of the detection results. The higher average precision indicates better detection performance.
- *Log-average Miss Rate* (MR^{-2}) is over False Positive Per Image (FPPI) ranging in $[10^{-2}, 10^{0}]$ (denoted as MR^{-2}). MR^{-2} is highly sensitive to false positives (FPs), especially FPs with high confidence scores significantly harm the MR^{-2} ratio. A lower MR^{-2} reflects the better performance.
- *Jaccard Index* (JI) represents the overlap between the detection results and the corresponding ground-truths [18], which is defined on the prediction sequence with decreasing confidence score. The detector with larger JI score tends to provide better performance.

B. Training Settings

We implement our proposed method in Pytorch with 2 NVIDIA GTX 1080Ti GPUs. For the CrowdHuman [18] dataset, we train FPN+, FPN++,Simple+ and Simple++ with a batch size of 6. The initial learning rate (LR) is set to 10^{-5}, which decays by a factor of 10 at 22^{nd} and 32^{nd} epoch, respectively. The training process stops at 50^{th} epoch. The momentum is set to 0.9 with a weight decay of 10^{-4}. All the networks are trained using an Adam optimizer. For CityPersons [50] dataset, we train CSP+ network with an initial LR is set to 2×10^{-4}, which decays by a factor of 10 at 110^{th} and 160^{th} epoch, respectively. The training stops at 200^{th} epoch with a total batch size of 4.

C. Experiments on the CrowdHuman Dataset

1) How to Choose the Most Appropriate and Approximate Annotation for Crowd Density Estimation Without any Extra Computational and Annotation Cost?

The CrowdHuman [18] dataset contains annotations for the head, the full-body, and the visible-body bounding box of pedestrian. Compared to the other annotations, the midpoint or the midpoint of the upper boundaries of the head bounding box is the most similar to the annotation required for the crowd density estimation. Intuitively, we should first consider the midpoints or the midpoints of the upper boundaries of all pedestrian head bounding boxes in the image as annotations for the crowd density estimation network. However, most pedestrian detection datasets exclude the pedestrian

(a) MFB (b) MH (c) MUH (d) MU

Fig. 4. We visualize four kinds of approximate crowd density estimation annotations. All the four approximate annotations are converted from existing pedestrian detection bounding boxes without any additional manual annotation cost. Specifically, MFB, MH, MUH, and MU represent the midpoint of full-body, head, upper boundary of the head bounding boxes and upper boundary of pedestrian bounding boxes, respectively, as shown in the dots with different colors.

TABLE II

OUR DDAD USES DIFFERENT CLASSES OF DENSITY ESTIMATION ANNOTATION, AND THE RESULTS OF PEDESTRIAN DETECTION SHOW THAT THE BEST PERFORMANCE OF PEDESTRIAN DETECTION IS OBTAINED WHEN WE USE THE MIDPOINT OF THE UPPER BOUNDARY OF THE PEDESTRIAN BOUNDING BOXES AS THE ANNOTATION

	Annotations	AP/%	MR $^{-2}$/%	JI/%
FPN [43]	–	86.65	42.43	79.49
FPN+	MFB	88.27	42.32	80.17
FPN+	MH	88.47	42.26	80.09
FPN+	MUH	**88.81**	42.16	**80.43**
FPN+	MU	88.71	**41.84**	80.41

head bounding box and crowd density annotations, and we require a high annotation cost to annotate them separately. We argue that the pedestrian detector should also pay more attention to the boundary of each pedestrian as well as the pedestrian's head region. Among the four boundaries (upper, lower, left, and right), the upper boundary contains more information and is visible even in heavily occluded scenes, which is a superiority over the other three boundaries. Therefore, we propose an alternative approach by considering the upper boundary of the midpoint of the pedestrian bounding box, which brings no extra annotation cost, as the annotation required for the crowd density estimator. This simple alternative approach has two advantages: (1) The annotation required for the crowd density estimation task is satisfied, and it is obtained from the existing pedestrian frames of the pedestrian detection dataset without any extra annotation cost; (2) It forces the shallow features to pay more attention to the upper boundary of the bounding boxes, which improves the pedestrian detection branch for small-scale pedestrians and reduces the error detection when there is no pedestrian. We also try to generate approximate annotations required for the crowd density estimation task based on the midpoint of the pedestrian full-body bounding boxes, which is also a cost-free method for generating annotations from the pedestrian datasets.

To validate our choice of the midpoint of the upper boundary on the pedestrian bounding box, we conduct extensive experiments to provide a comprehensive comparative analysis for other approximate crowd density annotations. The corresponding experimental results are shown in TABLE II, where MFB, MH, MUH, and MU represent the midpoint of full-body, head, upper boundary of the head, and upper boundary of pedestrian bounding boxes, respectively. Fig.4 illustrates the four kinds of approximate annotations for crowd density estimation. We keep all the training settings consistent except the kind of annotation for the crowd density estimation branch. It can be found that all our experiment results presented in TABLE II are better than the baseline FPN [42] results. We believe that all four kinds of approximate annotations we proposed improve the pedestrian detection performance, mainly due to the proper design of multi-task learning strategy. The crowd density estimation task forces backbone network to pay more attention to the pedestrian regions and increases the response values of these regions, which is beneficial for the pedestrian detection branch.

It can be observed from TABLE II that the two upper boundary-based approaches (MUH and MU) achieve better performance than the approaches based on the midpoint of the head and the full-body. Noted that the pedestrian detection task focuses on the regression of the pedestrian bounding boxes. We believe that exploiting meaningful upper boundary information enhances the attention of the pedestrian detection network to the upper boundary of the pedestrian frame. However, the midpoint of the full-body or the head is far away from the bounding box, yielding an insignificant performance gain. Moreover, the midpoint of the pedestrian frame tends to be occluded in crowded scenes, which also limits the performance gain in terms of pedestrian detection. The two ways for approximately obtaining crowd density estimation annotations based on pedestrian head frame (MH and MUH) outperform the way of utilizing the midpoint of the pedestrian frame owing to the proximity to the upper boundary and the virtual non-occlusion in crowded scenes. In particular, FPN+MUH achieves the best performance in terms of AP and JI evaluation metrics.

It should be noted that not all the pedestrian detection datasets are labeled with pedestrian head frames. We strive to find a new approximate alternatives method that meets the following three criteria: firstly, it should be close to the boundary; secondly, it should be visible even in heavily occluded scenes; and thirdly, no extra annotation should be required based on the existing pedestrian detection dataset. To meet these criteria, we propose utilizing the midpoint of the upper boundary of the pedestrian frame to approximate the annotation required for crowd density estimation. The midpoint of the upper boundary of the pedestrian frame is usually visible and also falls on the vertex of the head of the pedestrian. Therefore, applying

Fig. 5. Visual results of some example images in the CrowdHuman dataset. (a) Pedestrian detection results of FPN [42] (baseline). (b) Pedestrian detection results of our DDAD. (c) Pedestrian detection results of our IDAD. (d) Crowd density estimation result of our DDAD. (e) Crowd density estimation result of our IDAD. The red dashed boxes and ovals represent missed detection and missed detection areas, respectively. The solid red boxes represent false detection, while the solid green boxes represent accurate detection.

it as an approximate annotation for crowd density estimation can effectively facilitate backbone networks to focus on the upper boundary of the pedestrian. It can be observed from TABLE II that FPN+MU yields the best performance in terms of MR^{-2} metrics, and achieves an absolution gain of 0.59% compared to the baseline FPN [42]. Furthermore, it exceeds FPN+MUH by a margin of 0.32% in terms of MR^{-2} metrics. Meanwhile, FPN+MU is very competitive with FPN+MUH in terms of AP and JI evaluation metrics, being only 0.1% and 0.02% worse than FPN+MUH, respectively. Therefore, taking all these factors into account, it is reasonable to believe that the MU is the most suitable and effective annotation for crowd density estimation branch.

2) Collaborative of Two Pedestrian-Related Tasks

• Why crowd density estimation can help pedestrian detection? It is well known that multi-task learning can bring different kinds of noises to different tasks for more robust results and alleviate over-fitting. However, it may degrade the performance when designed randomly. In our proposed DDAD, the crowd density estimation features act as a noise for the pedestrian detection branch, which forces shallow backbone to pay more attention to the upper boundary of bounding boxes. Therefore, our DDAD reduces a lot of error detection boxes in the upper boundary of which with the low-density response, as shown in Fig.1 and Fig.5. Meanwhile, the crowd density estimation task focuses more on crowded areas, as well as small-scale pedestrian areas far away from the camera, which enhances the representation ability of related-area features in shallow layers. Therefore, it is natural to improve the pedestrian detection branch's ability to detect small-scale pedestrians and reduce the error detection, which increases the number of true positives (improvement in terms of MR^{-2}) and reduces the number of false detected bounding boxes, respectively. It makes a higher Recall and Precision, which increases the area of P-R curve (Improvement in terms of AP). As the experimental results shown in TABLE III, pedestrian detection task equipped with crowd density estimation can improve the former performance. Specifically, our DDAD (FPN+) gains 0.59% improvement in terms of MR^{-2}, as well as 2.06% and 0.92% in terms of AP and JI, respectively. What's more, our DDAD achieves an outstanding performance without any extra computational burdens compared to the baseline pedestrian detector in the inference phase.

• Effectiveness of deep feature fusion for pedestrian detection. As mentioned above, our DDAD improves pedestrian detection performance mainly because the two pedestrian-related tasks are collaborative. Besides, our DDAD works in a multi-task learning fashion such that pedestrian detection and crowd density estimation share the same backbone network to efficiently extract informative features. To further explore the collaboration of these two pedestrian-related tasks, we fused the high-level features of these two tasks in deep layers and then used for pedestrian detection, named IDAD. The corresponding results are reported in TABLE III. It can be found that our IDAD achieves a significant performance improvement over DDAD. In general, the larger the number of parameters, the better the model performance in the field of deep learning. To explore whether the performance improvement is simply due to the increased parameters of the model, we design an experiment without the supervision of crowd density estimation network (crowd density estimation related losses constantly equal to 0) based on FPN++, while keeping all other parameters same. According to the experiment results shown in TABLE III, the best MR^{-2} of FPN++ with crowd density estimation is 40.85%, which obtains an improvement of 0.69% than the FPN++ without

TABLE III

THE EXPERIMENTAL RESULTS OF PEDESTRIAN DETECTION BASED ON TWO BASELINES DETECTORS (FPN [42] AND SIMPLE OF CROWDDET [43]). THE FPS MEANS FRAMES PER SECOND AND THE CDE REPRESENTS THE SUPERVISION OF CROWD DENSITY ESTIMATION

Model	Paras(M)	FPS	CDE	AP/%	$MR^{-2}/\%$	JI/%
FPN	64.75	**12.34**		86.65	42.43	79.49
FPN+	64.75	**12.34**	✓	88.71	41.84	80.41
FPN++	75.63	7.49		89.09	41.54	80.46
FPN++	75.63	7.49	✓	**89.46**	**40.85**	**80.62**
Simple	64.76	12.18		89.97	41.67	82.25
Simple+	64.76	12.18	✓	91.70	41.24	82.63
Simple++	75.64	7.43		92.06	40.92	83.16
Simple++	75.64	7.43	✓	**92.58**	**39.70**	**83.58**

crowd density estimation. Therefore, it is effective to introduce the supervision of crowd density estimation during feature fusion. Furthermore, we visualize some challenging examples in the CrowdHuman [18] dataset, as shown in Fig.5, which shows that our DDAD and IDAD reduce the error detection rate and missed detection rate compared to the FPN [42] baseline.

• DDAD for Crowd Density estimation in crowded scenes. To optimize the network of crowd density estimation, we first propose to utilize the midpoint of the upper boundary of the pedestrian bounding boxes in an image as the annotation for crowd density estimation. In the CrowdHuman [18] dataset, the midpoint of the upper boundary of many pedestrian boundary boxes is close to a pedestrian head. However, it is natural that there will be some pedestrian bounding boxes whose midpoints of the upper boundary are not close to the pedestrian head, as many complex scenes exist in the dataset. Therefore, the annotations of the crowd density estimation generated by our method are not exactly equivalent to the annotations required for crowd density estimation and it is only an approximate equivalence. Moreover, due to the existence of the ignore region where the pedestrian group are difficult to annotate in the CrowdHuman [18] dataset, the total number of pedestrian bounding boxes in the training set and validation set are less than that of the pedestrians present in the images. Of course, this situation only happens to the images where the ignore region exists. With these two main constraints, can our method be proven to be a good density estimator, and how effective can it be?

To verify the effectiveness of our crowd density estimator, we show the estimation performance on the CrowdHuman [18] dataset in TABLE IV. On the CrowdHuman [18] dataset with an average of 22.64 pedestrians per image without crowd density estimation annotation, our method achieves a low MAE of 4.3. It can be observed in Fig.5 that our crowd density estimator performs well in the occluded scenes.

Furthermore, our IDAD fuses high-level feature maps from crowd density estimation with pedestrian detection in several deep layers to generate informative features only for pedestrian detection, which is shown in Fig.3. As the experiment results shown in TABLE IV, if the information flow is not blocked from pedestrian detection branch to the crowd density estimation branch in the back-propagation phase, the counting performance of our IDAD can be inferior to that of DDAD. Noted that deep layers contain rich semantic information, so simply fusing high-level features may deteriorate the overall performance if the mutual coordination between the two branches (i.e., pedestrian detection branch and crowd density estimation branch) remains absent. Intuitively, the deep layer features for classification and localization tasks in pedestrian detection may not be compatible with the high-level semantics features required for crowd density estimation. Therefore, the simply feature fusion may interfere with the high-level feature of crowd density estimation in the backpropagation phase, which degenerates its counting performance. To alleviate this problem, we block the high-level information flow from the pedestrian detection to the crowd density estimation branch during backpropagation, while retaining that from the latter to the former in our IDAD. In this case, the two pedestrian-related tasks share the same shallow backbone network, and the high-level information flow exists only in forward inference from crowd density estimation branch to pedestrian detection branch. The counting performance of crowd density estimation is as follows: MAE is 4.24, and MSE is 9.41, which exceeds that of FPN+ (DDAD). We believe that the better pedestrian detection features lead to better shallow feature representation and therefore improve the performance of crowd density estimation, which validates that these two tasks are complementary to each other. Also, the deep layer feature fusion promotes the performance improvement of pedestrian detection tasks. Finally, our DDAD and IDAD provide a baseline results for the subsequent crowd density estimation task in the CrowdHuman [18] dataset.

TABLE IV

THE DENSITY ESTIMATION ACHIEVED BY OUR DDAD ON THE CROWDHUMAN [18] DATASET IN TERMS OF MAE AND MSE EVALUATION METRICS. THE BLOCKED MEANS THAT WE FORBID THE GRADIENT OF THE PEDESTRIAN DETECTION BRANCH TO BE PROPAGATED TO THE CROWD DENSITY ESTIMATION BRANCH, WHEN THE GRADIENT IS BACK-PROPAGATED. IT SHOULD BE NOTED THAT FPN+ DOES NOT HAVE BLOCKED OPERATION

	Blocked	MAE	MSE
FPN+	-	4.31	9.72
FPN++	✗	4.34	9.97
FPN++	✓	4.24	9.41

TABLE V

COMPARATIVE RESULTS OF VARIOUS PEDESTRIAN DETECTION METHODS ON THE CROWDHUMAN [18] VALIDATION SET. HIGHER VALUES OF AP AND JI INDICATE BETTER PERFORMANCE, WHICH IS IN CONTRAST TO THE MR^{-2}

Methods	AP/%	MR $^{-2}$/%	JI/%
YOLO v5s [45]	85.50	49.15	-
FPN [43] baseline	86.65	42.43	79.49
Adaptive NMS [52]	84.71	49.73	–
JointDet [12]	–	46.50	–
Repulsion Loss [53]	85.64	45.69	–
S-RCNN [54]	90.7	44.7	81.4
D-DETR [55]	91.5	43.7	–
PEDR [56]	91.6	43.7	–
PBM [57]	89.29	43.35	–
CrowdDet [43](simple)	89.97	41.67	82.25
CrowdDet [43](refine)	90.30	41.28	**82.63**
Ours(YOLO v5s+)	86.00	48.68	-
Ours(FPN+)	88.71	41.84	80.41
Ours(Simple+)	**91.70**	**41.24**	**82.63**
Ours(FPN++)	89.46	40.85	80.62
Ours(Simple++)	**92.58**	**39.70**	**83.58**

3) Comparison With State-Of-The-Art

In this Section, we compare the performance of our proposed DDAD with several existing state-of-the-art pedestrian detectors. The comparative results are shown in TABLE V. It can be clearly observed that our method outperforms most of the pedestrian detectors and achieves significant improvement in comparison to our baseline approach. Specifically, compared to FPN [42] baseline pedestrian detector, our DDAD (FPN+) achieves an absolute gain of 0.59% in terms of MR^{-2} and still achieves the performance gain of 2.06% and 0.92% in terms of AP and JI, respectively. We also applied our method on the simple version of CrowdDet [43], our DDAD (Simple+) improves the AP and JI by 1.73% and 0.38%, respectively, while reducing the MR^{-2} by 0.43%, which is even better than the refined version of CrowdDet [43]. When it comes to the industrial application for pedestrian detection tasks, we also applied our DDAD on the YOLO v5s [45] baseline pedestrian detector. The performance of model Yolo v5s+ exceeds that of YOLO v5s in terms of both AP and MR^{-2} evaluation metrics. It should be noted that our proposed DDAD achieves state-of-the-art performance without any extra annotation cost and computation cost in the training phase and inference phase, respectively, when compared to our baseline pedestrian detectors. Similarly, our IDAD achieves a significant performance improvement compared to the baseline pedestrian detectors, which also demonstrates the effective of our proposed crowd density estimation assisted pedestrian detection. In our experiments, not only shallow feature sharing, but also high-level semantic feature fusion can improve the pedestrian detection performance. The major reason behind it is the correlation between these two pedestrian-related tasks. The crowd density estimation task focuses more on the regions with pedestrian head, which effectively reduces the missed pedestrian bounding boxes and suppresses the false detection bounding boxes where there are no pedestrians.

D. Experiments on CityPersons Dataset

We also conduct some experiments on CityPersons [50] dataset, and our method achieves comparative results with the existing state-of-the-art approaches as shown in TABLE VI. On the reasonable-occluded subset, in comparison to our baseline pedestrian detector CSP [44], our DDAD gains an improvement of 0.25% and 0.1% in terms of MR^{-2} with and without offset setting, respectively. We believe that there are few crowded scenarios in the CityPersons [50] dataset, and our DDAD achieves a slight performance improvement in the reasonable-occluded subset. However, on the heavily-occluded subset, we achieve a significant performance improvement based on our CSP [44] baseline pedestrian detector. In heavily occluded scenes, the pedestrian detection algorithm may predict some false detections and degrade performance. Our proposed multi-task learning strategy utilizes the crowded density estimation task, which is specifically designed for crowded scenes, to be trained together with the pedestrian detection task. It can effectively enhance the feature expressiveness at the upper boundary for the pedestrian detection task, reducing the number of false pedestrian prediction bounding boxes with the midpoint of which density response does not exist. Therefore, our proposed DDAD alleviates the impact of heavily occluded scenes on pedestrian detection performance and improves pedestrian detection performance.

TABLE VI

COMPARISONS OF DIFFERENT METHODS WITH THE MR^{-2} EVALUATION METRICS ON THE CITYPERSONS [50] VALIDATION SET. THE RESULTS ARE REPORTED ON THE ORIGINAL IMAGE SIZE (1024×2048 PIXELS)

Method	Reasonable(%)	Heavy(%)	Small(%)
Repulsion Loss [53]	13.22	56.85	–
OR-RCNN [7]	12.81	55.68	–
ALFNet [58]	12.01	51.90	19.00
CircleNet [59]	11.77	50.22	–
KGSNet [60]	10.96	39.68	–
CSP [44](w/o offset)	11.40	49.9	18.2
CSP [44](with offset)	11.00	49.3	16.0
CSP(w/o offset)+	11.31	37.85	16.51
CSP(with offset)+	**10.75**	**37.65**	**15.47**

V. CONCLUSION

In this paper, we propose a simple yet effective multi-task learning approach, namely Detachable crowd Density estimation Assisted pedestrian Detection (DDAD), which improves the pedestrian detection performance in crowded scenes. Specifically, we design a crowd density estimation branch that assists the pedestrian detection task for a better feature representation. During the network training phase, the extra crowd density estimation branch introduces a little computational overhead compared to the vanilla pedestrian detection networks when trained alone. However, our proposed DDAD achieves satisfactory performance without sacrificing inference speed as the proposed crowd density estimation branch is simply detached in the inference phase. Furthermore, we fuse the high-level features from crowd density estimation with pedestrian detection, resulting in an improved version of DDAD, denoted as IDAD, which demonstrates that the two pedestrian-related tasks in our multi-task learning framework collaborate with each other. In addition, our method can be

seamlessly applied to both single-stage and two-stage pedestrian detectors. To alleviate the lack of annotations required for crowd density estimation in the pedestrian detection datasets, we devise a cost-free method that approximates the midpoints of the upper boundaries of all pedestrian bounding boxes in the image as annotations for crowd density estimation. In our future work, we will extend our DDAD and IDAD to generic object detection tasks.

REFERENCES

[1] L. Chen et al., "Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3234–3246, Jun. 2021.

[2] Y. Zhang, Y. Zhang, and R. Su, "Pedestrian-safety-aware traffic light control strategy for urban traffic congestion alleviation," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 1, pp. 178–193, Jan. 2021.

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[4] X. Song, K. Zhao, W.-S. Chu, H. Zhang, and J. Guo, "Progressive refinement network for occluded pedestrian detection," in *Proc. ECCV*, 2020, pp. 32–48.

[5] J. Wu, C. Zhou, M. Yang, Q. Zhang, Y. Li, and J. Yuan, "Temporal-context enhanced detection of heavily occluded pedestrians," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13427–13436.

[6] J. Xie et al., "Count- and similarity-aware R-CNN for pedestrian detection," in *Proc. ECCV*, 2020, pp. 88–104.

[7] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Li, "Occlusion-aware R-CNN: Detecting pedestrians in a crowd," in *Proc. ECCV*, 2018, pp. 637–653.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[9] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Proc. NIPS*, 2014, pp. 1–9.

[10] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1751–1760.

[11] Y.-L. Hou and G. K. H. Pang, "People counting and human detection in a challenging situation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 1, pp. 24–33, Jan. 2011.

[12] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Relational learning for joint head and human detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10647–10654.

[13] K. Zhang, F. Xiong, P. Sun, L. Hu, B. Li, and G. Yu, "Double anchor R-CNN for human detection in a crowd," 2019, *arXiv:1909.09998*.

[14] C. Zhou and J. Yuan, "Bi-box regression for pedestrian detection and occlusion estimation," in *Proc. ECCV*, 2018, pp. 135–151.

[15] C.-Y. Lin, H.-X. Xie, and H. Zheng, "PedJointNet: Joint head-shoulder and full body deep network for pedestrian detection," *IEEE Access*, vol. 7, pp. 47687–47697, 2019.

[16] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020.

[17] Y. Jiao, H. Yao, and C. Xu, "PEN: Pose-embedding network for pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1150–1162, Mar. 2021.

[18] S. Shao et al., "CrowdHuman: A benchmark for detecting human in a crowd," 2018, *arXiv:1805.00123*.

[19] C. C. Loy, K. Chen, S. Gong, and T. Xiang, "Crowd counting and profiling: Methodology and evaluation," in *Modeling, Simulation and Visual Analysis of Crowds*. New York, NY, USA: Springer, 2013.

[20] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179–2195, Dec. 2009.

[21] B. Leibe, E. Seemann, and B. Schiele, "Pedestrian detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 878–885.

[22] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[23] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[24] A. B. Chan and N. Vasconcelos, "Bayesian Poisson regression for crowd counting," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 545–551.

[25] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2547–2554.

[26] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 833–841.

[27] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu, "CrowdNet: A deep convolutional network for dense crowd counting," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 640–644.

[28] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Fully convolutional crowd counting on highly congested scenes," 2016, *arXiv:1612.00220*.

[29] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 589–597.

[30] 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.

[31] X. Wang, C. Shen, H. Li, and S. Xu, "Human detection aided by deeply learned semantic masks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2663–2673, Aug. 2020.

[32] P. Kishore, S. Das, P. S. Mukherjee, and U. Bhattacharya, "ClueNet: A deep framework for occluded pedestrian pose estimation," in *Proc. BMVC*, 2019, p. 245.

[33] Y. Zhao, Z. Yuan, and B. Chen, "Accurate pedestrian detection by human pose regression," *IEEE Trans. Image Process.*, vol. 29, pp. 1591–1605, 2020.

[34] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Li, and X. Zou, "Pedhunter: Occlusion robust pedestrian detector in crowded scenes," in *Proc. AAAI*, 2020, pp. 10639–10646.

[35] 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. 4966–4974.

[36] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert, "Density-aware person detection and tracking in crowds," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2423–2430.

[37] J. Ren et al., "Accurate single stage detector using recurrent rolling convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 752–760.

[38] C. Shang, H. Ai, Z. Zhuang, L. Chen, and J. Xing, "ZoomNet: Deep aggregation learning for high-performance small pedestrian detection," in *Proc. Asian Conf. Mach. Learn.*, 2018, pp. 486–501.

[39] H. Zhang, K. Wang, Y. Tian, C. Gou, and F.-Y. Wang, "MFR-CNN: Incorporating multi-scale features and global information for traffic object detection," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8019–8030, Sep. 2018.

[40] S. Wang, J. Cheng, H. Liu, F. Wang, and H. Zhou, "Pedestrian detection via body part semantic and contextual information with DNN," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3148–3159, Nov. 2018.

[41] 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.

[42] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[43] 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. 12211–12220.

[44] 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. 5182–5191.

[45] G. Jocher et al., "Ultralytics/YOLOv5: V6.1—TensorRT, TensorFlow edge TPU and OpenVINO export and inference," Zenodo, Feb. 22, 2022, doi: 10.5281/zenodo.6222936.

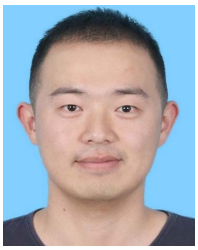
[46] V. K. Valloli and K. Mehta, "W-Net: Reinforced U-Net for density map estimation," 2019, *arXiv:1903.11249*.

[47] V.-S. Huynh, V.-H. Tran, and C.-C. Huang, "Iuml: Inception U-Net based multi-task learning for density level classification and crowd density estimation," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Oct. 2019, pp. 3019–3024.
[48] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[50] 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. 4457–4465.
[51] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[52] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6452–6461.
[53] 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.
[54] P. Sun et al., "Sparse R-CNN: End-to-end object detection with learnable proposals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14454–14463.
[55] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.
[56] M. Lin et al., "DETR for crowd pedestrian detection," 2020, *arXiv:2012.06785*.
[57] 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.
[58] 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)*, 2018, pp. 618–634.
[59] T. Zhang, Z. Han, H. Xu, B. Zhang, and Q. Ye, "CircleNet: Reciprocating feature adaptation for robust pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4593–4604, Nov. 2020.
[60] Y. Zhang, Y. Bai, M. Ding, S. Xu, and B. Ghanem, "KGSNet: Key-point-guided super-resolution network for pedestrian detection in the wild," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2251–2265, May 2021.

Wenxiao Tang (Student Member, IEEE) was born in Fujian, China, in 1996. He received the B.S. degree from the School of Electrical Engineering and Automation, College of Fuzhou University, Fuzhou, China, in 2019. He is currently pursuing the Ph.D. degree with the South China University of Technology, Guangzhou, China. His current research interests include pedestrian detection, pose estimation, and knowledge distillation.

Kun Liu received the master's degree from the School of Automation, Guangdong University of Technology, Guangzhou, China, in 2019. He is currently pursuing the Ph.D. degree with the South China University of Technology, Guangzhou. His research interests include computer vision, machine learning, object detection, and document analysis.

M. Saad Shakeel received the Ph.D. degree from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, in 2019. From 2019 to 2022, he worked as a Scientific Researcher at the School of Automation, Guangdong University of Petrochemical Technology, Maoming, China. He is currently working as a Post-Doctoral Research Fellow at the School of Automation Science and Engineering, South China University of Technology, Guangzhou, China. His research interests include computer vision, pattern recognition, and deep learning.



Hao Wang received the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2019. He is currently working as a Post-Doctoral Researcher at Guangdong Airport Baiyun Information Technology Company Ltd., and South China University of Technology. His research interests include image processing, pattern recognition, and computer vision.

Wenxiong Kang (Member, IEEE) received the M.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2003, and the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2009. He is currently a Professor with the School of Automation Science and Engineering, South China University of Technology. His research interests include biometrics identification, image processing, pattern recognition, and computer vision.