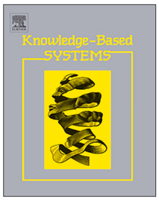
Knowledge-Based Systems 241 (2022) 108254

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Multi-expert learning for fusion of pedestrian detection bounding box

Zhi-Ri Tang[a], Ruihan Hu[b,*], Yanhua Chen[c,*], Zhao-Hui Sun[d], Ming Li[e]

[a] *School of Physics and Technology, Wuhan University, Wuhan, China*
[b] *Guangdong Key Laboratory of Modern Control Technology, Guangdong Institute of Intelligent Manufacturing, Guangzhou, China*
[c] *Department of Geography, The University of Hong Kong, Hong Kong, China*
[d] *Department of Industrial Engineering, Shanghai Jiao Tong University, Shanghai, China*
[e] *School of Intelligent Systems Science and Engineering, Jinan University (Zhuhai Campus), Zhuhai, China*

ARTICLE INFO

Article history:
Received 19 August 2021
Received in revised form 16 January 2022
Accepted 19 January 2022
Available online 31 January 2022

Keywords:
Pedestrian detection
Fine-tuning
Multi-expert learning
Fuse bounding boxes

ABSTRACT

Performance of pedestrian detection, which is one of the essential tasks in automatic drive, relies heavily on a large number of labels. Some researchers proposed unsupervised domain adaptive frameworks to improve the detection accuracy in wild datasets to reduce the need for labels. However, it is not a down-to-earth and cost-effective way for deploying these frameworks in practical engineering because it needs both source and target data for training. Unlike the former research, this work presents a new fine-tuning method without using source and target data for unsupervised detection. In this work, different well-trained models from the source domain are regarded as less-accurate experts in the wild domain, where a multi-expert learning algorithm is applied to learn from the difference between these models and fuse bounding boxes to present more accurate detection results. Experimental results on three common pedestrian detection datasets show that our method can efficiently improve the detection accuracy under unsupervised settings. Our method can also achieve better performance without source and target data involved compared with state-of-the-art works.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

Pedestrian detection, which is a task to get classification results and bounding boxes of pedestrians, is recently one of the most important research topics in computer vision. Pedestrian detection has been applied to many areas such as automatic drive technology [1–3] and video surveillance [4,5]. With the development of artificial intelligence, most pedestrian detection frameworks are developed based on deep learning methods. However, current detection frameworks have the same drawbacks as deep learning methods, which rely heavily on a large amount of annotated data for training. In addition, the generalization ability of a deep learning-based framework is not good enough, which means the performance will decrease a lot in the wild datasets (without labeled data for training). In other words, the generalization ability of well-trained pedestrian detection frameworks is still not feasible for practical applications.

In recent years, some works have been presented to improve the detection performance in wild datasets. One of the most comment methods is unsupervised domain adaptation [6–8], which bridges the gap between the source domain (with labels) and the target domain (without labels) from the distributions of two domains. However, it is usually an impractical and time-consuming way for engineering applications because it needs to bring all the source data, source labels, and target data to the engineering site for training. Another common way is to use a scene translation framework to generate a new target-like source domain [9, 10], which also needs to get the distributions and data from target data. Although the above works have achieved good performance based on cross-domain settings, an in-situ and more practical method to improve the generalization ability of well-trained models needs to be proposed.

Inspired by the above, an in-situ fine-tuning method for the well-trained pedestrian detection framework is proposed in this work. In general, detection results in wild datasets by well-trained models are inaccurate. Hence, a multi-expert learning method is applied, which regards well-trained models as inaccurate experts and can learn more precise information from the difference among well-trained models. Different from the traditional unsupervised domain adaptation method, which is shown in Fig. 1(a), both the source and target data are engaged in the training process. The flowchart of the proposed framework is shown in Fig. 1(b), where only the well-trained models from the source domain are needed in the fine-tuning process. The proposed method fine-tunes detection results in a wild dataset without using source and target data, which is a much easier and cheaper way to deploy in practical engineering. In addition,

[*] Corresponding authors.
E-mail addresses: rh.hu@giim.ac.cn (R. Hu), chen7whu@gmail.com (Y. Chen).

https://doi.org/10.1016/j.knosys.2022.108254
0950-7051/© 2022 Elsevier B.V. All rights reserved.

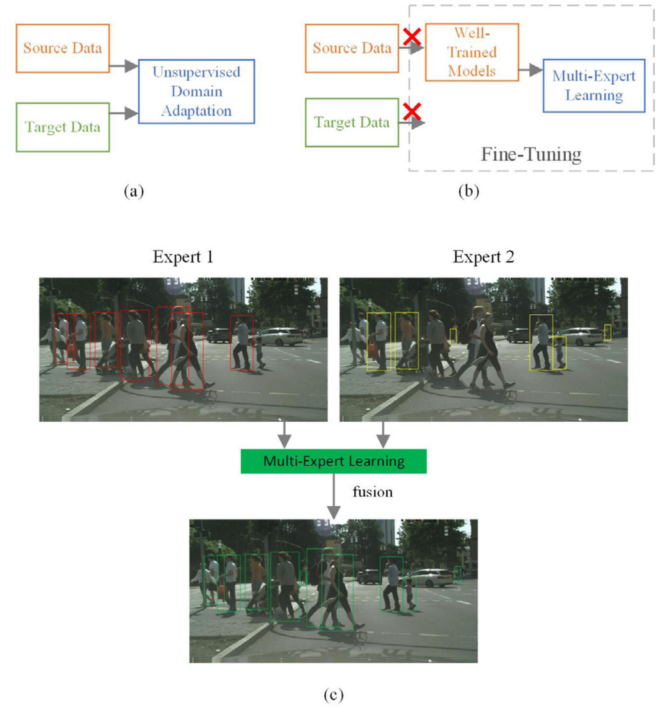


Fig. 1. (a) The flowchart of traditional unsupervised domain adaptation, where the source and target data are engaged in the training process. (b) The flowchart of the proposed multi-expert learning for fine-tuning process, in which only well-trained models are needed. (c) An overview of the proposed method, where two inaccurate models can be fused to better unsupervised detection performance.

an overview of the proposed framework is shown in Fig. 1(c), in which two well-trained models are regarded as two experts with different detection results in the target image. More accurate results can be obtained via a fusion process through the proposed method.

To evaluate the performance of the proposed method, three common pedestrian detection datasets, Caltech [11], KITTI [12] and CityPersons [13] are applied for experiment. Experimental results show that the proposed multi-expert learning for the fusion of bounding boxes can efficiently improve generalization under unsupervised settings. It can also give better performance than state-of-the-art works without using source and target data.

The main novelties of this paper can be summarized as follows: (1) To improve the generalization ability of well-trained models in wild datasets, a multi-expert learning method for fine-tuning is presented, which regards well-trained models as inaccurate experts and learns from the difference among these experts; (2) Different from other recent works on unsupervised domain adaptation, the proposed method does not need source and target data for extra training, which means it presents an easier and cheaper way to deploy in practical engineering; (3) Experiments on three common pedestrian detection datasets verify the efficiency of the proposed multi-expert learning for fine-tuning, which is better than other state-of-the-art works.

The rest of this work is arranged as follows: Section 2 presents a review of state-of-the-art works. Section 3 gives the proposed method, and Section 4 shows the experimental results. Finally, the conclusion and discussion on the future works are introduced in Section 5.

2. Related works

2.1. Object detection

In recent years, deep learning has achieved great success in many areas including computer vision [14–16], image processing [17–20], and signal processing [21–24]. The performance on many traditional computer vision tasks such as object detection [25,26] is improved by deep learning frameworks. As a specific task of object detection, many popular deep learning frameworks on object detection can also be deployed for pedestrian detection [27,28], which contains two main tasks, including classification and localization.

In general, recent mainstream pedestrian detection frameworks can be divided into two main parts: anchor-based and anchor-free framework [29]. Anchor-based frameworks, which use a large number of preset anchors to get detection results for pedestrians. Some classic anchor-based frameworks include Faster R-CNN [30] and YOLO [31] while some novel anchor-free detectors emerge in recent years such as FCOS [32] and CSP [33].

2.2. Unsupervised pedestrian detection

The performance of the pedestrian detector, which is similar to the performance of deep learning frameworks, relies heavily on a large number of labels. When faced with wild datasets, the generalization ability of well-trained models will decrease a lot. Many feasible ways have been introduced in recent years, which can be divided into four main types:

(1) Scene-specific detection framework: Some researchers proposed a multi-feature learning framework for better representation and generalization ability [34]. In addition, some others also presented a four-step scene-specific detection framework based on feature transfer [4]. Although the above works can effectively improve the detection performance in the target domain, the design of a scene-specific detector always involves complex steps, limiting the real-time requirements and practical applications.

(2) Unsupervised domain adaptation: To solve this issue, some works applied domain adaptation [35,36] to bridge the gap between source and target domains, which helps the entire framework to learn the general feature rather than build a dataset-specific framework. Inspired by the domain adaptation methods of other computer vision tasks, some presented Maximum Mean Discrepancy (MMD) [37] loss in unsupervised domain adaptation pedestrian detection [38]. Furthermore, some other researchers also introduced a self-paced domain adaptation method with progressive latent models [39]. In conclusion, the basis of domain adaptation is access to source and target data. A domain adaptation framework with good performance is obtained through joint training with source and target feature information. However, for some practical applications, it is not a feasible way to bring millions of data and annotations to the target domain. In addition, for some mobile terminals, a well-trained model with good generalization ability, rather than a complex training process using source and target data in mobile equipment, is often more suitable.

(3) Learning from noisy labels: For the wild datasets without annotations, a feasible way is to generate pseudo labels using well-trained models. Some proposed an automatic adaptation detection framework, which is trained with the high-confidence pseudo labels [40]. However, the performance is limited because the accuracy of pseudo labels will influence the retraining a lot. In other words, the performance heavily depends on the well-trained models from other annotated domains. Furthermore, for the object detection task, the accuracy of pseudo labels includes the classification accuracy and the localization of bounding boxes.

Although a slight offset of the box does not influence the test performance in the source domain, it will lead to a wrong training direction in the wild datasets.

(4) Data augmentation using GAN: With the development of GAN in recent years, many works applied GAN for data augmentation to bridge the gap between the source and target domains. One feasible way is to apply GAN frameworks to generate a new dataset [10], which has the source data and labels with distributions of the target domain and can also be regarded as an image-level adaptation. Another way is to generate more pedestrians and add them to source data randomly [41], which can provide more features to train a detector with better generalization ability. However, the above works need both the source and target data during the entire training process, which is not easy to deploy and does not meet the actual engineering requirements to some extent.

Although many previous works introduced some ways to improve the unsupervised detection performance in the unlabeled wild datasets, there is still no work mentioned how to build a learning system to fine-tune the well-trained models in the wild datasets and help achieve a good unsupervised performance. Compared with the current works, the fine-tuning way needs no source and target data during the training process, which provides a more accessible and cheaper method for deploying unsupervised pedestrian detection in this work.

2.3. Multi-expert learning

Multi-expert learning [42] is presented for learning from multiple noisy annotators. It has many applications on medical diagnose because annotations from different doctors are usually with different confidence scores [43]. For other real-world applications, many less-accurate labels also exist due to some non-expert annotators [44,45]. From above, it is a feasible way to adopt multi-expert learning to present better results from well-trained models because these models can be regarded as noisy annotators in wild datasets. However, applying multi-expert learning in unsupervised pedestrian detection is still a challenge. Hence, we introduce a new way, which regards the well-trained models from the source domain as experts with different experiences. Furthermore, the feasibility and limitations of this learning system are also presented in this work.

3. Method

The flowchart of the proposed method is shown in Fig. 2. First, an anchor-free detection backbone named CSP is adopted as the detector backbone in this work. Similar to other pedestrian detection frameworks, the CSP detector has a feature extraction module and then gives detection results. In this work, two well-trained CSP detection models from the source domain give inaccurate inference results for the wild datasets, which include detection categories and bounding boxes. Then the bounding boxes given from two CSP models are divided into three levels: Level 0 represents the overlap part of different models is bigger than a fixed threshold so that the bounding boxes can be regarded as "true" label in target domain; Level 2 represents the overlap part of different models is very small, and all the bounding boxes will be kept; For Level 1, which gives inaccurate inference results, the multi-expert learning method is applied to fuse the bounding boxes from different models. Through the above process, a fine-tuning method aiming to give better unsupervised detection results in the wild datasets is presented. Experimental results on a number of benchmark datasets and comparisons with state-of-the-art works verify the efficiency of the proposed framework.

3.1. Backbone detector

In this work, a state-of-the-art pedestrian detector, CSP [33], is used as the backbone detection framework. Different from traditional anchor-based detectors such as Faster R-CNN [30] and YOLO [31], CSP is a novel anchor-free detection framework, which uses a detection head to give the center and scale of detection targets. Generally speaking, anchor-based detection frameworks always present a series of anchor boxes with different sizes and scales in the test images. Then, the classification results of each anchor boxes are presented, where only the boxes with higher classification accuracy are kept. There is no doubt that the performance of anchor-based frameworks depends heavily on the preset boxes. In other words, when faced with a new scene or dataset, the parameters of preset boxes always need to be modified to ensure detection performance. On the contrary, the anchor-free detector adopted in our work can present the detection results directly. Without a series of preset boxes, the performance of an anchor-free detector is robust when facing different scenes and datasets.

The framework of the CSP detector is shown in Fig. 3. First, a concatenated feature map from a four-level feature pyramid is obtained. Following the general settings of feature pyramid [25], Stage 2, 3, 4, and 5 adopt 1/4, 1/8, 1/16, and 1/32 downsampling, respectively. Then, an L2 Norm is adopted for each stage to rescale them to 10. Following the setting of other work [46], the scale of concatenated feature map is set to $H/4$ and $W/4$ of the original input image. Apart from Stage 2, whose size is the same as the setting of the concatenated feature map, dilated convolutions are adopted to the other three stages to keep their output. Then, one detection head, which has one 3×3 convolutional layer and two 1×1 convolutional layers, is applied to obtain the centers and scales from the concatenated feature map. Furthermore, a Gaussian mask $G(.)$ is adopted to obtain the center point via regression as follows:

$$Mask_{ij} = \max_{n=1,\ldots,N} G(i, j; x_n, y_n, \delta_{w_n}, \delta_{h_n}) \tag{1}$$

where the Gaussian function is formulated as:

$$G(i, j; x_n, y_n, \delta_{w_n}, \delta_{h_n}) = e^{-(\frac{(i-x)^2}{2\delta_w^2} + \frac{(j-y)^2}{2\delta_h^2})} \tag{2}$$

where N is the number of targets in the scene and center point is localized by (x_n, y_n, w_n, h_n). In addition, δ_{w_n} and δ_{h_n} are proportional to the height and width of the target.

Furthermore, normal Cross-Entropy loss is not feasible for detecting the center points because only a few points in the images belong to the positive class. All other pixels in the images are in the negative class (background). Hence, although the classification of center points and background is a binary classification problem, it also faces a very serious class imbalance issue. To address the extreme class imbalance problems between pedestrians and background, the focal loss [26] is adopted as follows:

$$L_c = -\frac{1}{N} \sum_{i=1}^{W/r} \sum_{j=1}^{H/r} \alpha_{ij} (1 - p_{ij})^{\gamma} \log(p_{ij}) \tag{3}$$

which has

$$p_{ij} = \begin{cases} p_{ij} & y_{ij} = 1 \\ 1 - p_{ij} & otherwise \end{cases} \tag{4}$$

$$\alpha_{ij} = \begin{cases} 1 & y_{ij} = 1 \\ (1 - M_{ij})^{\beta} & otherwise \end{cases} \tag{5}$$

The probability of the target pedestrian in location (i, j) is $p_{ij} \in [0, 1]$. The annotations for pedestrians are $y_{ij} \in \{0, 1\}$.

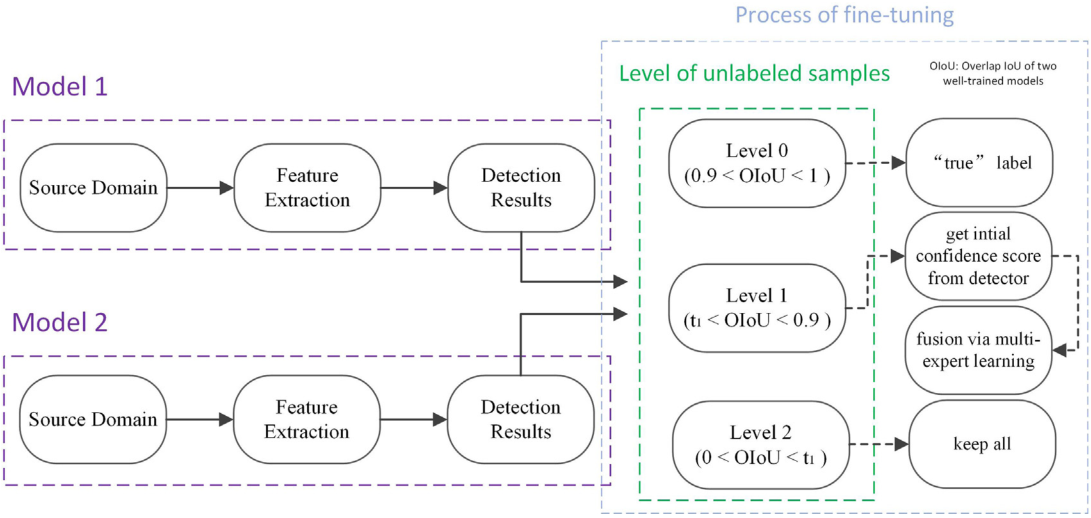

Process of fine-tuning
Level of unlabeled samples
OIoU: Overlap IoU of two well-trained models
Model 1
Source Domain
Feature Extraction
Detection Results
Model 2
Source Domain
Feature Extraction
Detection Results
Level 0 (0.9 < OIoU < 1)
Level 1 (t₁ < OIoU < 0.9)
Level 2 (0 < OIoU < t₁)
"true" label
get intial confidence score from detector
fusion via multi-expert learning
keep all

Fig. 2. The flowchart of the proposed method. First, inaccurate detection results in the wild dataset are obtained via well-trained models from the source data. Because a certain degree of bounding box offset can also be considered correct detection, some boxes with a high overlap area could be regarded as the "true" label for one target. In addition, if the overlap area is too low, they should be considered as boxes for different targets due to some crowd humans in scenes. Hence, all the bounding boxes can be divided into three different levels based on the overlap area of different models: (1) Level 0 is regarded as true labels; (2) Level 1 needs the multi-expert learning method to fuse the boxes; and (3) Level 2 will keep all the corresponding bounding boxes.

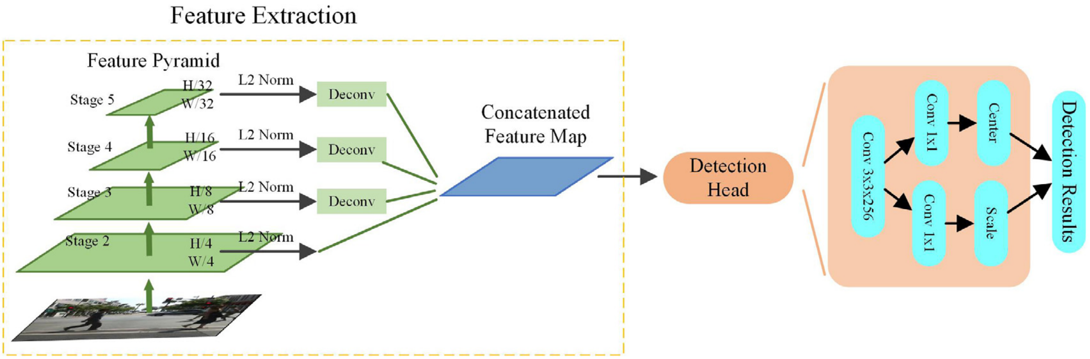


Fig. 3. The framework of the CSP detector. First, a feature pyramid is adopted to extract the features from the original images and then is concatenated into one feature map. Furthermore, a detection head with two main branches for obtaining the center and scale of targets are presented.

Based on the above focal loss, the weights of easy samples in the background can be decreased to some degree, and the classifier could give better performance on center points and some hard negative points.

In addition, the scale regression aims to measure the distance between groundtruth and predicting labels, which is calculated through a smooth L1 loss as follows:

$$L_s = \frac{1}{N}\sum_{n=1}^{N} SmoothL1(s_n, gt_n) \tag{6}$$

where s_n and gt_n are output of the framework and groundtruth of pedestrians, respectively.

Apart from the loss functions for center and scale, another smooth L1 loss is also adopted to form an offset loss to fine-tune the bounding boxes of targets and improve the detection performance:

$$\begin{aligned} L_o = {} & \frac{1}{N}\sum_{n=1}^{N} SmoothL1(x_n/r, x_n^*/r) \\ & + \frac{1}{N}\sum_{n=1}^{N} SmoothL1(y_n/r, y_n^*/r) \end{aligned} \tag{7}$$

where x_n/r and y_n/r are also outputs of the framework, and x_n^*/r and y_n^*/r are the true labels.

Finally, the entire loss function of the CSP detector is a weighted sum of the above three losses as follows:

$$L = \lambda_c L_c + \lambda_s L_s + \lambda_o L_o \tag{8}$$

where λ_c, λ_s, and λ_o are weights for center, scale prediction, and offset losses, respectively.

3.2. Maximum likelihood estimation

Set i be the number of test instances, where N is the total quantity. Let j be the number of well-trained models (experts). After the inference process, each model can give a set of bounding boxes for the test image. For a specific instance, let $OIoU$ be the overlap area of bounding boxes from j experts and y^j be the inference results from jth model. When the overlap area is larger than 90% of the bounding box given by j^th model, the inference results can be regarded as "true" label:

$$\hat{y}_i^j = \begin{cases} 1 & , \quad OIoU_i^j \geq 0.9 * IoU_i^j \\ 0 & , \quad otherwise \end{cases} \tag{9}$$

If the bounding box is true ($y = 1$), the true positive rate can be formulated as

$$\alpha_i^j = P(\hat{y}_i^j = 1 | y = 1) \tag{10}$$

For the bounding box is not true, the 1 - false positive rate can be formulated as

$$\beta_i^j = P(\hat{y}_i^j = 0|y = 0) \tag{11}$$

Based on the setting, the experts give a fair judgment on each instance without any bias. In other words, the difference between instances has nothing to do with α_i^j and β_i^j.

Due to a mass of bounding boxes in some specific images/scenes, it is very common that bounding boxes from several experts for one instance have overlapped with that for another instance. To judge whether the bounding boxes given by the jth expert is for ith instance or not, a function is presented as:

$$\hat{y}_i^j = \begin{cases} \hat{y}_i^j & , \quad OIoU_i^j \geq t_1 * IoU_i^j \\ 0 & , \quad otherwise \end{cases} \tag{12}$$

where 0 represents the "default" of the bounding box from the jth expert. The threshold t_1 is set for different scenes with different densities of instance.

To judge whether the inference results from experts are correct or not, a linear discriminator is proposed as:

$$\hat{y}_i^j = \begin{cases} 1 & , \quad w^T x \geq t_2 \\ 0 & , \quad otherwise \end{cases} \tag{13}$$

where $w, x \in \mathbb{R}^d$. $f(x) = w^T x$ is linear classification function and the degree of tightness of this discriminator depends on the threshold t_2. Hence the true positive rate can be rewritten as

$$P(\hat{y}_i^j = 1|w, x) = \sigma(w^T x) \tag{14}$$

where $\sigma(w^T x)$ is a function for logistic regression of discrimination as

$$\sigma(w^T x) = \frac{1}{1 + e^{-w^T x}} \tag{15}$$

The actual gold standard of maximum estimation for each instance is unknown. Hence, Expectation-Maximization (EM) algorithm is used to solve this maximization problem with hidden data. The training data D and parameters Ω from N test instance and E experts can be formulated as

$$D = \{x_i, y_i^1, y_i^2, \ldots, y_i^E\}_{i=1}^N \tag{16}$$

$$\Omega = \{w, \alpha, \beta\} \tag{17}$$

which have $i \in [1, N], j \in [1, E]$, and

$$\alpha = [\alpha^1, \alpha^2, \ldots, \alpha^E] \tag{18}$$

$$\beta = [\beta^1, \beta^2, \ldots, \beta^E] \tag{19}$$

Hence, the maximum likelihood of Ω can be factored as

$$\hat{\Omega}_{ML} = \{\hat{w}, \hat{\alpha}, \hat{\beta}\} = \arg\max_{\Omega}\{\ln[P(D|\Omega)]\} \tag{20}$$

Because the distributions of training samples in D are independent, the likelihood $P(D|\Omega)$ can be formulated as

$$\begin{aligned} P(D|\Omega) &= \prod_{i=1}^N P(y_i^1, y_i^2, \ldots, y_i^E|x_i, \Omega) \\ &= \prod_{i=1}^N (P(y_i^1, y_i^2, \ldots, y_i^E|y_i = 1, \alpha)P(y_i = 1|x_i, w) \\ &+ P(y_i^1, y_i^2, \ldots, y_i^E|y_i = 0, \beta)P(y_i = 0|x_i, w)) \end{aligned} \tag{21}$$

As mentioned before, each expert gives the inference results (bounding boxes) independently. So we have

$$\begin{aligned} P(y_i^1, y_i^2, \ldots, y_i^E|y_i = 1, \alpha) &= \prod_{j=1}^E P(y_i^j|y_i = 1, \alpha^j) \\ &= \prod_{j=1}^E (\alpha^j)^{y_i^j}(1 - \alpha^j)^{1-y_i^j} \end{aligned} \tag{22}$$

$$\begin{aligned} P(y_i^1, y_i^2, \ldots, y_i^E|y_i = 0, \beta) &= \prod_{j=1}^E P(y_i^j|y_i = 0, \beta^j) \\ &= \prod_{j=1}^E (\beta^j)^{1-y_i^j}(1 - \beta^j)^{y_i^j} \end{aligned} \tag{23}$$

So (21) can be formulated as

$$P(D|\Omega) = \prod_{i=1}^N (m_i^j p_i + n_i^j(1 - p_i)) \tag{24}$$

which have

$$p_i := \sigma(w^T x_i) \tag{25}$$

$$m_i^j := P(y_i^j|y_i = 1, \alpha_i^j) \tag{26}$$

$$n_i^j := P(y_i^j|y_i = 0, \beta_i^j) \tag{27}$$

Take the hidden data y into (21), we can get the log-likelihood as

$$\begin{aligned} \log P(D, y|\Omega) = \sum_{i=1}^N \sum_{j=1}^E y_i \log m_i^j p_i \\ + (1 - y_i) \log n_i^j(1 - p_i) \end{aligned} \tag{28}$$

3.3. Expectation-maximization optimization

In E-step of EM algorithm, the expectation of log-likelihood is got first. Then the maximization process of lower bounds of log-likelihood is carried out in M-step. The expectation of log-likelihood is

$$\begin{aligned} E[\log P(D, y|\Omega)] = \sum_{i=1}^N \sum_{j=1}^E \mu_i^j \log m_i^j p_i \\ + (1 - \mu_i^j) \log n_i^j(1 - p_i) \end{aligned} \tag{29}$$

which have

$$\mu_i^j = P(y_i = 1|y_i^j, x_i, \Omega) \tag{30}$$

Based on Bayes' theorem, it has

$$\begin{aligned} \mu_i^j &\propto P(y_i^j|y_i = 1, \Omega)P(y_i = 1|x_i, \Omega) \\ &= \frac{m_i^j p_i}{m_i^j p_i + n_i^j(1 - p_i)} \end{aligned} \tag{31}$$

In M-step, the estimation of parameters Ω can be got from the gradient of (29), which have

$$\alpha_i^j = \frac{\sum_{i=1}^N \mu_i^j y_i^j}{\sum_{i=1}^N \mu_i^j} \tag{32}$$

$$\beta_i^j = \frac{\sum_{i=1}^N (1 - \mu_i^j)(1 - y_i^j)}{\sum_{i=1}^N (1 - \mu_i^j)} \tag{33}$$

Table 1
The comparisons among three pedestrian detection datasets.

Dataset	Caltech [11]	KITTI [12]	CityPersons [13]
Number of training image	42,782	7,481	2,975
Number of test image	4,024	7518	500
Resolution ratio	640 × 480	1224 × 370	2048 × 1024
Country	1	1	3
City	1	1	18
Season	1	1	3
Unique person	1,273	6,336	19,654

Provided some experts are more reliable than others, the priors of the specific experts need to be set by the inference performance of each expert:

$$\alpha^j_{prior} = TP^j_{IoU=50} \tag{34}$$

$$\beta^j_{prior} = 1 - FP^j_{IoU=50} \tag{35}$$

According to the above EM algorithm, a soft variable of each expert for each instance can be got from it:

$$v^j_i = \frac{\mu^j_i}{\sum_{j=1}^{E} \mu^j_i} \tag{36}$$

Let $\Theta = \{x, y, w, h\}$ be the variables of bounding boxes, k be generated bounding box, and G be the total number bounding boxes in one scene, which has $k \in [1, G]$. Suppose there have E_a experts given an original inference results for the ith instance, which has $E_a \in [1, E]$. Hence the bounding boxes after fine-tuning can be calculated by

$$\Theta_k := \frac{\sum_{j=1}^{E_a} v^j_i \Theta^j_i}{E_a} \tag{37}$$

where the Θ^j_i is given by original inference results from the jth expert. For the generated bounding boxes, the confidence score can be presented as

$$\lambda_{\Theta_k} = \frac{\sum_{j=1}^{E_a} v^j_i}{E_a} \tag{38}$$

According to (12), a non-maximum suppression (NMS) is applied to select the generated bounding boxes to avoid numerous inference results for the same instance. For random two bounding boxes Θ_{k_p} and Θ_{k_q} with $k_p, k_q \in [1, G]$, the NMS is processed as

$$\Theta_{k_p} = \begin{cases} \Theta_{k_p} & , \quad OIoU_{\Theta_{k_p} \cap \Theta_{k_q}} < t_1 * IoU_{\Theta_{k_p}} \\ \Theta_{k_p} & , \quad OIoU_{\Theta_{k_p} \cap \Theta_{k_q}} \geq t_1 * IoU_{\Theta_{k_p}} \\ & \quad \& \quad \lambda_{\Theta_{k_p}} \geq \lambda_{\Theta_{k_p}} \\ \Theta_{k_q} & , \quad OIoU_{\Theta_{k_p} \cap \Theta_{k_q}} \geq t_1 * IoU_{\Theta_{k_p}} \\ & \quad \& \quad \lambda_{\Theta_{k_p}} < \lambda_{\Theta_{k_p}} \end{cases} \tag{39}$$

where the $\lambda_{\Theta_{k_p}}$ and $\lambda_{\Theta_{k_q}}$ are the confidence scores of two bounding boxes given by (38).

4. Experiment

4.1. Experimental setup

Three benchmark pedestrian detection datasets, including Caltech [11], KITTI [12], and CityPersons [13], are applied to verify the performance of the proposed multi-expert learning method, where the number of images and resolution ratio of three datasets are presented in Table 1. Experiments in pedestrian detection focus on cross-domain detection to evaluate how the proposed method can improve the unsupervised detection performance.

Caltech [11], which includes 42,782 and 4024 images in the training and test sets, respectively, is got from the vehicle-mounted video. There are over 2300 pedestrians with annotations in the entire dataset. To better improve and evaluate the unsupervised detection performance, a new and more accurate set of annotations [27] are used in the training and inference process. Some samples from Caltech are shown in Fig. 4(a).

KITTI [12] dataset has 7481 training images with high resolution. The training set is randomly divided into two subsets for training and test because the official test set has no public annotations, where the division rule follows [13]. Some samples from KITTI are shown in Fig. 4(b).

CityPersons [13] is a new pedestrian detection dataset that extracts from Cityscapes [47] dataset. There are a total of 2975 images got from 18 cities in the training set and 500 images from 3 cities in the validation set. Unlike Caltech and KITTI datasets, images in the CityPersons dataset have much higher resolution and more pedestrians. Some samples from CityPersons are shown in Fig. 4(c).

Intuitively, CityPersons is a ‘harder’ dataset because it has relatively complex distributions and features. The comparison of training subsets in three different datasets is presented in Table 1. From the table, the CityPersons dataset is got from a more complex environment, including more countries, cities, and seasons. Further, although the Caltech dataset has over 2300 pedestrians in the entire dataset, only 1273 unique persons are in the training subset. For the CityPersons dataset, it is a total of 19,654 unique persons, which means it has about 7 persons per image.

We choose the above three datasets for experiments because there is a great distance between distributions of them. In other words, there will have a huge gap between cross-domain detection and fully-supervised learning performance in Caltech, KITTI, and CityPersons. Hence, these three datasets are suitable for verifying the proposed method’s performance. Evaluation follows the standard metric [11] of Caltech dataset, which is log-average Miss Rate over False Positive Per Image (FPPI) (also denoted MR^{-2}) under $IoU = 0.5$.

In addition, the detection framework is implemented following CSP detector [33], which uses Tensorflow and Keras to build the overall framework in one RTX 2080Ti GPU. Batch size for training is set as 1. Hyper-parameters α and β in Focal loss of CSP detector are set as 2 and 4, respectively, according to [26]. t_1 is set as 0.3 based on experimental setup in [33], while t_2 is set as 0.5 according to [42]. Furthermore, ResNet-50 [48] is used as backbone for the detection framework.

4.2. Experimental results

The fusion results of the proposed multi-expert learning method are shown in Fig. 5. Fig. 5(a) and (b) are detection results from two different well-trained models in Caltech dataset. It can be seen that Fig. 5(a) only gives 6 bounding boxes for pedestrians in this image and misses many other pedestrians. Fig. 5(b) presents a lot of bounding boxes, but many of them are wrong. The above detection results from two models are common inaccurate cross-domain detection conditions, which are not suitable for distributions of the target domain and overly sensitive to the feature of the target scene, respectively.

Fig. 5(c) presents fusion results of the proposed multi-expert learning methods without NMS process based on Fig. 5(a) and (b). Fig. 5(d) presents the fusion results with NMS. From the fused bounding boxes, it can be seen that the proposed method can efficiently learn from inaccurate experts and give relatively accurate detection results.

For different levels in the process of fine-tuning, the overview of fusion is shown in Fig. 6. In Fig. 6(a), which is shown for Level

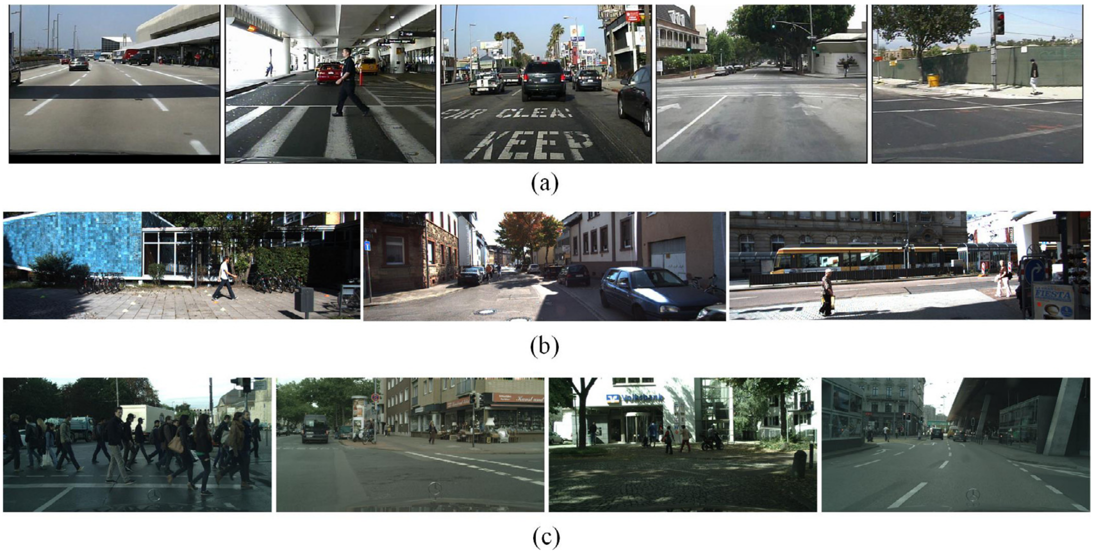


Fig. 4. Some samples from dataset (a) Caltech, (b) KITTI, and (c) CityPersons. It can be seen that the Caltech dataset is with relatively low resolution, while KITTI and CityPersons are with high resolution. Furthermore, the CityPersons dataset has more pedestrians in one image than the other two datasets.

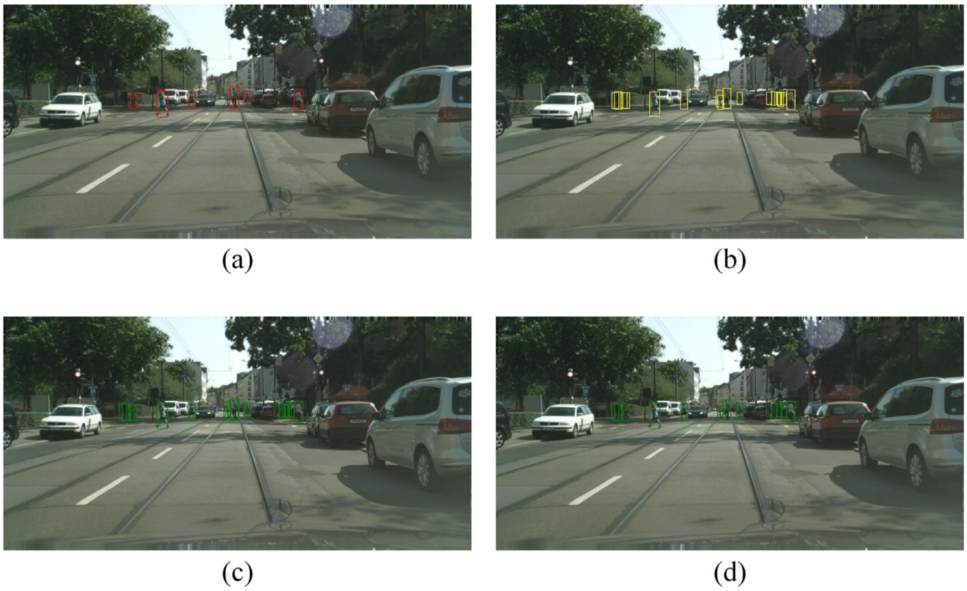


Fig. 5. Fusion results of the proposed multi-expert learning in the CityPersons dataset. (a) and (b) give the detection results with bounding boxes from two well-trained models from the Caltech dataset. (c) presents fusion results without NMS process while (d) proposes fusion results with NMS.

0, the overlap part from different experts is big enough so that the bounding boxes given by both experts can be regarded as detection results with high confidence scores. In Fig. 6(b), some inaccurate bounding boxes are taken for examples, where the fusion process can efficiently improve the detection accuracy for these pedestrians. In Fig. 6(c), some samples which are only given by one expert are shown. It can solve some missing detection problems from the difference between experts.

The miss rate curve in Caltech is also shown in Fig. 7. The blue curve is unsupervised detection results in Caltech from well-trained models in CityPersons using CSP detector, while the orange curve gives unsupervised results using our method. The lower curve shows better detection results for the miss rate curve in detection tasks. From the figure, our method can help to improve the unsupervised detection performance efficiently.

To better evaluate the performance of the proposed method, qualitative analysis for different datasets is shown in Table 2. The results by CSP [33] show the cross-domain detection results from the CSP detector without any modifications. Furthermore, the oracle results give the supervised detection results by CSP detector for comparisons. For MR^{-2}(%), lower numbers give better performance. Hence from Table 2, the proposed method help to improve unsupervised detection performance by about 9%–10% in Cal2City (abbreviation for Caltech → CityPersons), 4%–5% in Cal2KIT (abbreviation for Caltech → KITTI), and 2%–3% in City2Cal (abbreviation for CityPersons → Caltech). The above results show that our method can efficiently improve the unsupervised detection performance. Specifically, our method presents the best improvement compared with other settings for the Cal2City, which has the biggest gap between supervised and unsupervised detection.

4.3. Comparisons with state-of-the-art works

To better show the merits of the proposed multi-expert learning method, comparisons with state-of-the-art works in City2Cal

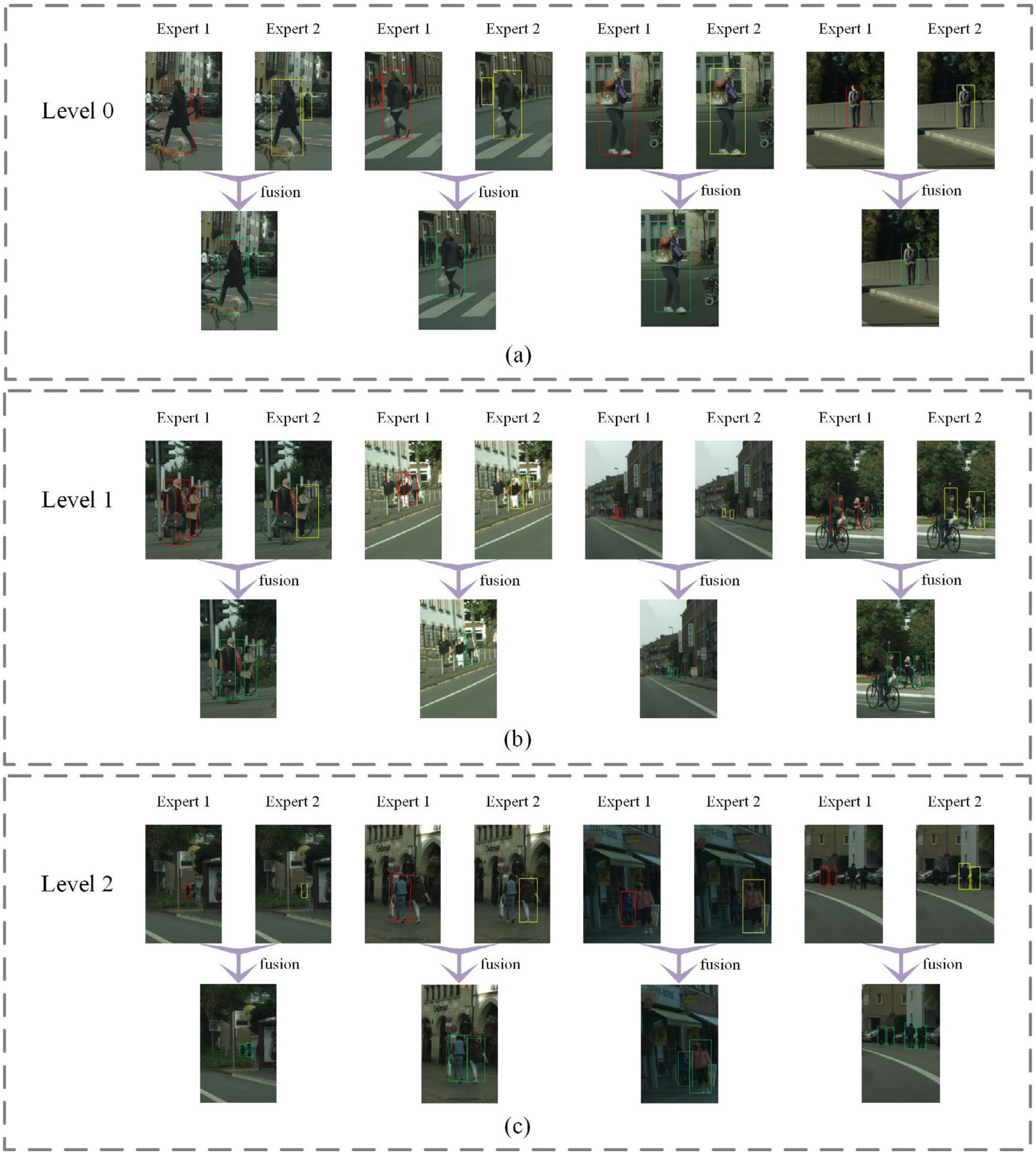


Fig. 6. Fusion results for different levels in process of fine-tuning: (a) Level 0; (b) Level 1; and (c) Level 2.

Table 2
Cross domain detection results using the proposed method. The optimal experimental results are shown in bold.

Caltech → CityPersons		CityPersons → Caltech		Caltech → KITTI	
Method	MR^{-2}(%)	Method	MR^{-2}(%)	Method	MR^{-2}(%)
CSP [33]	53.98	CSP [33]	21.64	CSP [33]	12.81
Ours	**44.38**	Ours	**16.95**	Ours	**10.67**
Oracle	11.0	Oracle	4.54	Oracle	7.86

are given in Table 3. ACF [49], CSP [33], Adapted FasterRCNN [13], and ALFNet [50] are single detector which proposed for achieving accurate detection performance in a specific pedestrian detection dataset. In other words, these frameworks do not need source/target data when testing in the target domain. Compared with the above four frameworks in City2Cal, our method can succeed by about 34%–35% for ACF, 4%–5% for CSP and Adapted FasterRCNN, and 8%–9% for ALFNet.

Apart from the above detectors, two state-of-the-art works focusing on improving the generalization ability of pedestrian detection are chosen for comparisons. PRNet [51] presented a progressive refinement network for detecting occluded pedestrians. In addition, alterable receptive fields are presented in PRNet for detecting small size or parts targets. Experimental results show that this framework can efficiently improve the unsupervised cross-domain pedestrian detection performance in City2Cal, which also needs no source/target data when deploying the entire framework. Our method can also give better generalization under the same settings compared with this work. APGAN [41] presented an intuitive way to improve the generalization ability of

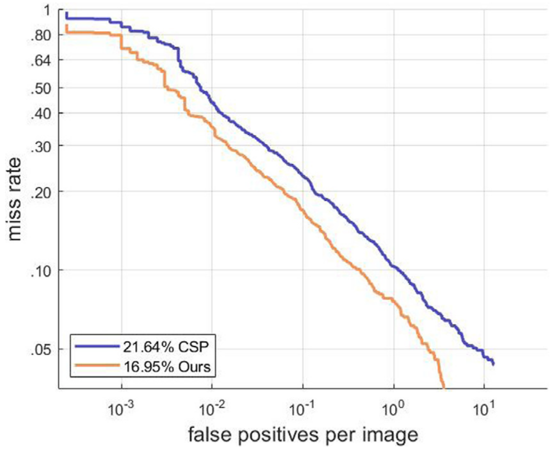


Fig. 7. The miss rate curve before and after the proposed entire method from CityPersons to Caltech (The lower curve shows better results).

Table 3
Unsupervised cross domain detection performance on City2Cal and comparisons with state-of-the-art works. The optimal experimental results are shown in bold.

CityPersons → Caltech		
Method	Use source/Target data	MR^{-2}(%)
ACF [49]	N/A	51.28
CSP [33]	N/A	21.64
Adapted FasterRCNN [13]	N/A	21.18
ALFNet [50]	N/A	25.0
PRNet [51]	N	18.3
APGAN [41]	Y	20.5
Ours	**N**	**16.95**
Oracle	N/A	4.54

Table 4
Unsupervised cross domain detection performance on Cal2City and comparisons with state-of-the-art works. The optimal experimental results are shown in bold.

Caltech → CityPersons		
Method	Use source/Target data	MR^{-2}(%)
ACF [49]	N/A	72.89
Adapted FasterRCNN [13]	N/A	46.91
CSP [33]	N/A	53.98
Ours	**N**	**38.42**
Oracle	N/A	11.0

the pre-trained models, which generated many new pedestrians in source data with similar distributions of the target domain. Although it is a feasible way to improve the unsupervised cross-domain detection performance, APGAN also needs source and target domains to help learn the distributions of different domains and generate new pedestrians. Compared with APGAN, our method can succeed by about 3%–4% without source and target data participation.

Furthermore, for two other settings, including Cal2City and Cal2KIT, the comparisons with the latest works are also included in Tables 4 and 5. For Cal2City, the proposed method can achieve the best performance without using source/target data. Specifically, compared with the second-best framework, CSP [33], our method can succeed by about 15%–16% in MR^{-2}(%). For Cal2KIT, although the gap between performance using CSP original detector and oracle results is very small, our method can also achieve the second-best performance with only 0.17% behind the best results.

Table 5
Unsupervised cross domain detection performance on Cal2KIT and comparisons with state-of-the-art works. The optimal experimental results are shown in bold.

Caltech → KITTI		
Method	Use source/Target data	MR^{-2}(%)
ACF [49]	N/A	49.99
Adapted FasterRCNN [13]	N/A	**10.50**
CSP [33]	N/A	12.81
Ours	**N**	10.67
Oracle	N/A	7.86

5. Conclusion

In this work, a multi-expert learning method is proposed to fuse pedestrian detection bounding box, which presents a feasible way to improve the generalization ability and unsupervised detection performance of well-trained models in the wild dataset. All the bounding boxes given by well-trained models from the source domain are divided into three levels and fused differently. Three benchmark pedestrian detection datasets are used to evaluate the performance of the proposed method. Experimental results show that our method can help to improve the unsupervised detection by about 2%–10%. Furthermore, compared with state-of-the-art works, the proposed can also achieve better performance by about 3%–4%. Even compared with the latest works using source or target data, our work can also achieve better performance.

Based on the experimental results, our work can effectively improve the generalization ability of current pedestrian detection models. In other words, the proposed framework is helpful for an automatic driving system to detect pedestrians in unknown scenes, which can improve the safety of autonomous driving. In addition, compared with previous works, including domain adaptation, retraining with noisy labels, etc., the proposed framework presents an in-situ and lightweight solution for an automatic driving system.

The main limitations of this work include: (1) The detection performance of the proposed framework depends on each expert to some degree. If the original experts can only give bad detection results in the wild datasets, two or more experts will not perform well. Besides, the performance of each expert also depends on many factors, such as the selection of models, source data, number of annotations, extracted features, etc. (2) Some empirical parameters are presented in this work, such as the threshold to determine the level of unlabeled samples during multi-expert fusion and the threshold of annotations' confidence score. There may be some changes in these parameters' settings for specific applications.

For future works, several directions can be focused on: (1) Although three benchmark pedestrian detection datasets extracted from real-world scenes are adopted, more datasets and more detection tasks should be explored to expand the applications of the proposed framework. (2) Although the proposed framework has achieved a good performance on some datasets, it is also essential to explore the empirical parameters and develop a more general learning system for autonomous driving. (3) Although the proposed framework presents an in-situ and lightweight solution and can improve the efficiency compared with previous works, it is also very hard to meet the real-time requirements of autonomous driving at high speed. Hence, how to propose a learning system to meet the real-time requirements and improve the generalization ability effectively might be a valuable direction in the future.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is partly supported by National Natural Science Foundation of China under Grant 52005218, in part by the National Natural Science Foundation of China Youth Fund under Grant 6210023461, in part by the Guangdong Academy of Sciences' (GDAS') Project of Science and Technology Development, China under Grant 2017GDASCX-0115 and Grant 2018GDASCX-0115, in part by the Guangdong Academy of Science for the Special Fund of Introducing Doctoral Talent, China under Grant 2021GDASYL-20210103087, in part by the Opening Foundation of Xinjiang Production and Construction Corps Key Laboratory of Modern Agricultural Machinery under Grant BTNJ2021003, in part by the Foundation of Engineering Research Center of Intelligence Perception and Autonomous Control, China, Ministry of Education of China (K100052021008).

References

[1] N. Ragesh, R. Rajesh, Pedestrian detection in automotive safety: understanding state-of-the-art, IEEE Access 7 (2019) 47864–47890.

[2] G. Shen, Z.-R. Tang, P. Shen, Y. Yu, HQ-trans: A high-quality screening based image translation framework for unsupervised cross-domain pedestrian detection, in: International Conference on Image and Graphics, Springer, 2021, pp. 16–27.

[3] G. Shen, Y. Yu, Z.-R. Tang, H. Chen, Z. Zhou, HQA-trans: An end-to-end high-quality-awareness image translation framework for unsupervised cross-domain pedestrian detection, IET Comput. Vis. (2021).

[4] X. Wang, M. Wang, W. Li, Scene-specific pedestrian detection for static video surveillance, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2013) 361–374.

[5] M. Bilal, A. Khan, M.U.K. Khan, C.-M. Kyung, A low-complexity pedestrian detection framework for smart video surveillance systems, IEEE Trans. Circuits Syst. Video Technol. 27 (10) (2016) 2260–2273.

[6] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: International Conference on Machine Learning, PMLR, 2015, pp. 1180–1189.

[7] K. Saito, K. Watanabe, Y. Ushiku, T. Harada, Maximum classifier discrepancy for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3723–3732.

[8] M. Long, H. Zhu, J. Wang, M.I. Jordan, Unsupervised domain adaptation with residual transfer networks, in: Adv. Neural Inf. Process. Syst., 2016, pp. 136–144.

[9] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, K. Kim, Image to image translation for domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4500–4509.

[10] H.-K. Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H.-Y. Tseng, M. Singh, M.-H. Yang, Progressive domain adaptation for object detection, in: The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 749–757.

[11] P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: A benchmark, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 304–311.

[12] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The kitti dataset, Int. J. Robot. Res. 32 (11) (2013) 1231–1237.

[13] S. Zhang, R. Benenson, B. Schiele, Citypersons: A diverse dataset for pedestrian detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3213–3221.

[14] Z. Tang, Z.-H. Sun, E.Q. Wu, C.-F. Wei, D. Ming, S. Chen, MRCG: A MRI retrieval system with convolutional and graph neural networks for secure and private iomt, IEEE J. Biomed. Health Inf. (2021).

[15] R. Zhu, Z. Tang, S. Ye, Q. Huang, L. Guo, S. Chang, Memristor-based image enhancement: High efficiency and robustness, IEEE Trans. Electron Devices 68 (2) (2020) 602–609.

[16] Z.-R. Tang, Q.-Q. Chen, Z.-H. Sun, P. Xiong, B.-H. Zhang, L. Jiang, E.Q. Wu, Few-sample generation of amount in figures for financial multi-bill scene based on GAN, IEEE Trans. Comput. Soc. Syst. (2021).

[17] Z. Tang, R. Zhu, R. Hu, Y. Chen, E.Q. Wu, H. Wang, J. He, Q. Huang, S. Chang, A multilayer neural network merging image preprocessing and pattern recognition by integrating diffusion and drift memristors, IEEE Trans. Cogn. Dev. Syst. (2020).

[18] R. Zhu, S. Ye, Z. Tang, P. Lin, Q. Huang, H. Wang, J. He, S. Chang, Influence of compact Memristors' stability on machine learning, IEEE Access 7 (2019) 47472–47478.

[19] Z. Tang, Y. Chen, Z. Wang, R. Hu, E.Q. Wu, Non-spike timing-dependent plasticity learning mechanism for memristive neural networks, Appl. Intell. 51 (6) (2021) 3684–3695.

[20] Z. Tang, Y. Chen, S. Ye, R. Hu, H. Wang, J. He, Q. Huang, S. Chang, Fully memristive spiking-neuron learning framework and its applications on pattern recognition and edge detection, Neurocomputing 403 (2020) 80–87.

[21] R. Hu, S. Zhou, Y. Liu, Z. Tang, Margin-based Pareto ensemble pruning: An ensemble pruning algorithm that learns to search optimized ensembles, Comput. Intell. Neurosci. 2019 (2019).

[22] R. Hu, Q. Mo, Y. Xie, Y. Xu, J. Chen, Y. Yang, H. Zhou, Z.-R. Tang, E.Q. Wu, AVMSN: An audio-visual two stream crowd counting framework under low-quality conditions, IEEE Access (2021).

[23] R. Hu, Z.-R. Tang, X. Song, J. Luo, E.Q. Wu, S. Chang, Ensemble echo network with deep architecture for time-series modeling, Neural Comput. Appl. (2020) 1–14.

[24] R. Hu, S. Zhou, Z.R. Tang, S. Chang, Q. Huang, Y. Liu, W. Han, E.Q. Wu, DMMAN: A two-stage audio–visual fusion framework for sound separation and event localization, Neural Netw. 133 (2020) 229–239.

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[27] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, How far are we from solving pedestrian detection? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1259–1267.

[28] S. Zhang, R. Benenson, B. Schiele, et al., Filtered channel features for pedestrian detection, in: CVPR, 1, (2) 2015, p. 4.

[29] S. Zhang, C. Chi, Y. Yao, Z. Lei, S.Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9759–9768.

[30] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Adv. Neural Inf. Process. Syst., 2015, pp. 91–99.

[31] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[32] Z. Tian, C. Shen, H. Chen, T. He, Fcos: Fully convolutional one-stage object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9627–9636.

[33] W. Liu, S. Liao, W. Ren, W. Hu, Y. Yu, High-level semantic feature detection: A new perspective for pedestrian detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5187–5196.

[34] X. Zeng, W. Ouyang, M. Wang, X. Wang, Deep learning of scene-specific classifier for pedestrian detection, in: European Conference on Computer Vision, Springer, 2014, pp. 472–487.

[35] D. Vazquez, A.M. Lopez, J. Marin, D. Ponsa, D. Geronimo, Virtual and real world adaptation for pedestrian detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (4) (2013) 797–809.

[36] L. Liu, W. Lin, L. Wu, Y. Yu, M.Y. Yang, Unsupervised deep domain adaptation for pedestrian detection, in: European Conference on Computer Vision, Springer, 2016, pp. 676–691.

[37] W. Liu, J. Li, B. Liu, W. Guan, Y. Zhou, C. Xu, Unified cross-domain classification via geometric and statistical adaptations, Pattern Recognit. 110 (2021) 107658.

[38] L. Liu, W. Lin, L. Wu, Y. Yu, M.Y. Yang, Unsupervised deep domain adaptation for pedestrian detection, in: European Conference on Computer Vision, Springer, 2016, pp. 676–691.

[39] Q. Ye, T. Zhang, W. Ke, Q. Qiu, J. Chen, G. Sapiro, B. Zhang, Self-learning scene-specific pedestrian detectors using a progressive latent model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 509–518.

[40] A. RoyChowdhury, P. Chakrabarty, A. Singh, S. Jin, H. Jiang, L. Cao, E. Learned-Miller, Automatic adaptation of object detectors to new domains using self-training, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 780–790.

[41] S. Liu, H. Guo, J.-G. Hu, X. Zhao, C. Zhao, T. Wang, Y. Zhu, J. Wang, M. Tang, A novel data augmentation scheme for pedestrian detection with attribute preserving GAN, Neurocomputing (2020).

[42] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, L. Moy, Learning from crowds., J. Mach. Learn. Res. 11 (4) (2010).

[43] Y. Duan, O. Wu, Learning with auxiliary less-noisy labels, IEEE Trans. Neural Netw. Learn. Syst. 28 (7) (2016) 1716–1721.

[44] K. Ma, X. Liu, Y. Fang, E.P. Simoncelli, Blind image quality assessment by learning from multiple annotators, in: 2019 IEEE International Conference on Image Processing, ICIP, IEEE, 2019, pp. 2344–2348.

[45] D. Zhou, S. Basu, Y. Mao, J. Platt, Learning from the wisdom of crowds by minimax entropy, Adv. Neural Inf. Process. Syst. 25 (2012) 2195–2203.

[46] T. Song, L. Sun, D. Xie, H. Sun, S. Pu, Small-scale pedestrian detection based on topological line localization and temporal feature aggregation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 536–551.

[47] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.

[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[49] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2014) 1532–1545.

[50] W. Liu, S. Liao, W. Hu, X. Liang, X. Chen, Learning efficient single-stage pedestrian detectors by asymptotic localization fitting, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 618–634.

[51] X. Song, K. Zhao, W.-S.C.H. Zhang, J. Guo, Progressive refinement network for occluded pedestrian detection, in: Proc. European Conference on Computer Vision, Vol. 7, 2020, p. 9.