SAN: Selective Alignment Network for Cross-Domain Pedestrian Detection

Yifan Jiao, Hantao Yao, *Member, IEEE*, and Changsheng Xu, *Fellow, IEEE*

***Abstract*— Cross-domain pedestrian detection, which has been attracting much attention, assumes that the training and test images are drawn from different data distributions. Existing methods focus on aligning the descriptions of whole candidate instances between source and target domains. Since there exists a giant visual difference among the candidate instances, aligning whole candidate instances between two domains cannot overcome the inter-instance difference. Compared with aligning the whole candidate instances, we consider that aligning each type of instances separately is a more reasonable manner. Therefore, we propose a novel Selective Alignment Network for cross-domain pedestrian detection, which consists of three components: a Base Detector, an Image-Level Adaptation Network, and an Instance-Level Adaptation Network. The Image-Level Adaptation Network and Instance-Level Adaptation Network can be regarded as the global-level and local-level alignments, respectively. Similar to the Faster R-CNN, the Base Detector, which is composed of a Feature module, an RPN module and a Detection module, is used to infer a robust pedestrian detector with the annotated source data. Once obtaining the image description extracted by the Feature module, the Image-Level Adaptation Network is proposed to align the image description with an adversarial domain classifier. Given the candidate proposals generated by the RPN module, the Instance-Level Adaptation Network firstly clusters the source candidate proposals into several groups according to their visual features, and thus generates the pseudo label for each candidate proposal. After generating the pseudo labels, we align the source and target domains by maximizing and minimizing the discrepancy between the prediction of two classifiers iteratively. Extensive evaluations on several benchmarks demonstrate the effectiveness of the proposed approach for cross-domain pedestrian detection.**

***Index Terms*— Cross-domain pedestrian detection, instance-level adaptation network, image-level adaptation network, pedestrian detection.**

I. Introduction

PEDESTRIAN detection that aims to predict a series of bounding boxes enclosing pedestrians for a given image, as a particular branch of general object detection, has been attracting more and more interests in both academia and industry. Driven by the surge of convolutional neural networks (CNN) [1], many CNN-based pedestrian detection approaches have been proposed to boost the performance [2]–[12]. However, these methods all assume that the training and test images have the same distribution, limiting the generalization of the proposed methods. As shown in Figure 1, using the detector inferred on the Caltech dataset [13] obtains a worse detection result on the CityPersons dataset [14].

To improve the generalization of the proposed methods on the unseen target domain, the critical factor is how to reduce the domain gap between the source domain and target domain, *e.g.,* domain adaptation [15]–[20] has been widely studied in the classification task. Inspired by domain adaptation, a lot of methods have been proposed for cross-domain detection [21]–[25]. These methods reduce the domain bias from the following two aspects: image-level domain adaptation, and instance-level domain adaptation, which can be regarded as the global-level and local-level domain alignments, respectively. For example, Cai *et al.* [24] embed the Faster R-CNN into the Mean Teacher framework with object relations to tackle the instance-level domain adaptation. Different from [24], Chen *et al.* [21] design two domain adaptation components to alleviate the domain discrepancy at both the image and instance levels. Although the above methods can reduce much domain bias between the source and target domains, they do not consider the instance alignment or simply consider the whole candidate instances during instance-level domain adaptation. In conclusion, they all pay a little attention to the instance-level domain adaptation. Due to the complexity of instance appearance, the candidate instances belonging to the same image have a large inter-instance difference. As a consequence, ignoring the effect of inter-instance difference will degrade the cross-domain detection performance.

As mentioned above, the instance-level domain adaptation plays a crucial role in cross-domain pedestrian detection. To reduce the effect caused by inter-instance difference, we consider that aligning the candidate proposals having similar distributions between source and target domains has

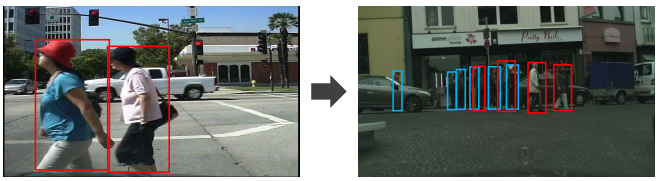
Manuscript received June 3, 2020; revised November 12, 2020 and December 27, 2020; accepted December 28, 2020. Date of publication January 20, 2021; date of current version January 27, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102205, in part by the National Natural Science Foundation of China under Grant 61902399, Grant 61721004, Grant U1836220, Grant U1705262, Grant 61832002, and Grant 61720106006, in part by the Beijing Natural Science Foundation under Grant L201001, and in part by the Key Research Program of Frontier Sciences, CAS, under Grant QYZDJ-SSW-JSC039. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sos S. Agaian. *(Corresponding author: Changsheng Xu.)*

Yifan Jiao is with School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail:yifanjiao1227@gmail.com).

Hantao Yao is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: hantao.yao@nlpr.ia.ac.cn).

Changsheng Xu is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: csxu@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TIP.2021.3049948



(a) Caltech (b) CityPersons

Fig. 1. Samples from the Caltech (a) and CityPersons (b) dataset. Since there exists an obvious visual difference between two datasets, merely applying the detector trained on the Caltech dataset generates many false detection results on the CityPersons, *e.g.,* red and blue bounding boxes represent the positive and negative results.

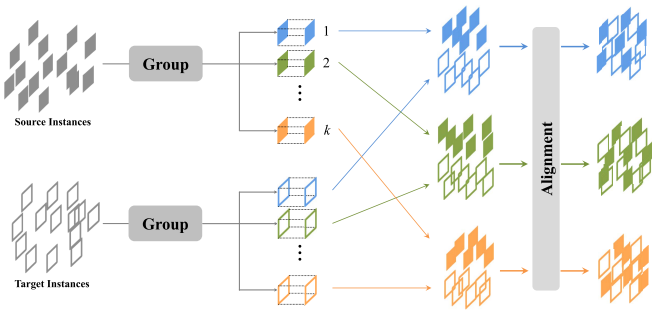


Fig. 2. The proposed method firstly clusters the candidate proposals into k groups for both source and target domains, and then apply the instance-level domain adaptation on the corresponding groups in the source and target domains.

less domain bias than aligning whole candidate proposals. As shown in Figure 2, we can firstly cluster the given candidate proposals into k groups for both source and target domains, and then apply the instance-level domain adaptation on the corresponding groups in the source and target domains. The advantage of the group-based instance-level domain adaptation is that it can overcome the inter-instance difference.

To achieve the above goals, we propose a novel Selective Alignment Network (SAN) for cross-domain pedestrian detection. As illustrated in Figure 3, the proposed network consists of three sub-networks: a Base Detector, an Image-Level Adaptation Network, and an Instance-Level Adaptation Network. Given an image, the Base Detector aims to predict the location of pedestrians. Similar to the Faster R-CNN, the Base Detector is composed of a Feature module, an RPN module and a Detection module. The Feature module is used to extract the feature map of the whole image, and the RPN module is applied to generate several candidate proposals used for prediction. To reduce the global visual bias, the Image-Level Adaptation Network is proposed to align the whole image feature with an adversarial domain classifier. Based on the candidate proposals generated by the RPN module for source and target domains, the Instance-Level Adaptation Network is proposed to align those candidate proposals between two domains in two steps. Firstly, the candidate proposals are gathered into several clusters by the Group module according to their visual feature for the source domain, and the pseudo label is further obtained for each candidate proposal according to the clustering results. After generating the pseudo labels, we propose an Alignment module to align the source and target domains by iteratively maximizing and minimizing the discrepancy between the prediction of two classifiers.

The main contributions can be summarized as follows:

(1) We propose a novel Selective Alignment Network to address the domain bias between the source and target domains for cross-domain pedestrian detection.

(2) We demonstrate that using the group-based instance alignment is an effective way to align the source and target domains for instance-level domain adaptation.

(3) The extensive experiments on three datasets, *i.e.,* Caltech [13], CityPersons [14] and COCOPersons [26], prove the effectiveness of the proposed model.

II. Related Work

In this section, we give a brief review about pedestrian detection, followed by the domain adaptation. After that, we give much discussion about cross-domain detection.

A. Pedestrian Detection

As a specific sub-task of object detection, pedestrian detection is the first and critical technology for many real-world applications. Traditional pedestrian detectors, such as ACF [27], LDCF [28] and Checkerboards [29], extend the Viola and Jones paradigm [30] to exploit various filters on Integral Channel Features (ICF) [31] with the sliding window strategy to localize each pedestrian. Recently, deep CNNs have been widely adopted for pedestrian detection and achieved the good performance [32]–[39]. CNN-based detectors can be roughly divided into two categories: the anchor-based approach, and the anchor-free approach. The anchor-based detector is performed by classifying and regressing anchor boxes with predefined scales and aspect ratios. The anchor-based methods can be treated as a two-stage approach or one-stage approach. The two-stage approach comprises separate proposal generation followed by confidence computation of proposals. For instance, Pang *et al.* [36] introduce a mask-guided attention network to emphasize on the visible pedestrian regions while suppress the occluded ones by modulating full body features. Zhou *et al.* [37] propose a discriminative feature transformation to handle occlusions for pedestrian detection. Different from the two-stage approach, the one-stage approach combines the proposal generation and classification as a single-stage regression. For example, AFLNet [33] proposes an asymptotic localization fitting strategy to evolve the default anchor boxes step by step into accurate detection. Lin *et al.* [34] focus on discriminative representation learning based on the original SSD architecture. The anchor-free approach bypassing the requirement of anchor boxes can detect pedestrians directly from an image. For example, CSP [38] not only demonstrates that a single center point is feasible for pedestrian localization, but also can generate bounding boxes with the scale prediction. Although great progress has been made, existing approaches all rely on the annotated training images, limiting their generalization.

B. Domain Adaptation

Domain adaptation, which adapts a model to a new domain without re-training from scratch, has been attracting intensive

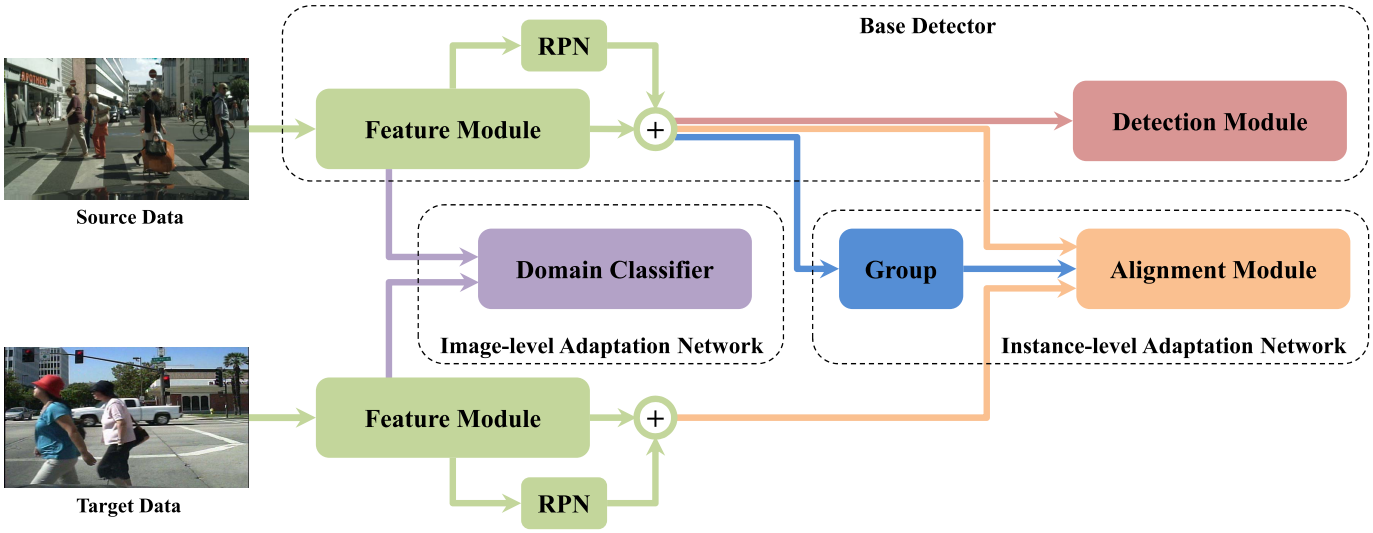


Fig. 3. The overall pipeline of the proposed model, which consists of three main parts: a Base Detector, an Image-Level Adaptation Network, and an Instance-Level Adaptation Network. The Base Detector is composed of a Feature module, an RPN module, and a Detection module. The Image-Level Adaptation includes a domain classifier. The Instance-Level Adaptation Network is made up of a Group module and an Alignment module. Notably, the parameters are shared between the modules with the same color.

interests recently. A typical approach is to estimate the domain gap formalized in certain ways and minimize it [40]–[42]. Furthermore, some recent methods [15]–[20], [43]–[45] propose the more effective ways to reduce the domain gap. For example, the GRL [15] proposes a rather trivial gradient reversal layer for domain adaptation. Mean Teacher [17] makes a faster feedback loop between the student and the teacher models. MCD [18] introduces a new approach to align distributions of source and target by utilizing the task-specific decision boundaries. MMEN [19] proposes a category discriminator, which classifies source samples accurately but is confused about the categories of target samples, to align the source and target samples.

C. Cross-Domain Detection

Inspired by the traditional domain adaptation, a lot of methods have been proposed for cross-domain detection [21]–[25], [46]–[49]. For example, Khodabandeh *et al.* [46] address the domain adaptation problem from the perspective of robust learning by formulating the problem as training with noisy labels. Based on the H-divergence theory, Chen *et al.* [21] design two domain adaptation components to reduce the domain discrepancy for image-level and instance-level simultaneously. Wang *et al.* [23] introduce a pairing mechanism over the source and target features to alleviate the issue of insufficient target domain samples, and further propose a bi-level module to adapt the trained source detector to the target domain. As mentioned, most existing methods can reduce much domain bias between the source and target domains, but they do not consider the instance alignment or simply consider all candidate instances during instance-level domain adaptation. Therefore, they all pay a little attention to the instance-level domain adaptation.

Among all existing methods, the most related work is [22], which focuses on mining and aligning the discriminative regions. However, there are two main differences with ours. Firstly, they mine the discriminative regions according to the location information of the candidate proposal, while we use the visual description. Since the core of domain adaptation is how to project the source and target samples into a common feature space, using the visual descriptions to align the instances between two domains is more robust than using the location information. Secondly, they design a weighting estimator to indicate how well a target region matches the source to establish the relationship between two domains, while our proposed method applies the instance-level domain adaptation on the corresponding groups in the source and target domains.

III. Methodology

A. Problem Formulation

Cross-domain pedestrian detection aims to recognize pedestrians belonging to the unseen target domain with the help of labeled source domain. Formally, the source data are defined as $\mathcal{D}_s = \{\mathcal{X}_s, \mathcal{B}_s\}$, where $\mathcal{X}_s$ are the source images, and $\mathcal{B}_s$ is the ground-truth bounding boxes for each pedestrian in $\mathcal{X}_s$. The $\mathcal{D}_t = \{\mathcal{X}_t\}$ is the target dataset, where $\mathcal{X}_t$ are the target images.

Since the source data $\mathcal{D}_s$ and the target data $\mathcal{D}_t$ have an obvious domain gap, *e.g.,* the pedestrians belonging to the source images and target images have an apparent visual difference, as shown in Figure 1. The critical problem of cross-domain pedestrian detection is how to reduce the domain gap between source and target domains. Unlike the image domain adaptation, which focuses on reducing the domain bias of the global-level image description, the cross-domain pedestrian detection contains two types of domain bias: global visual bias, and local visual bias. The global visual bias, which denotes the image-level visual difference between the

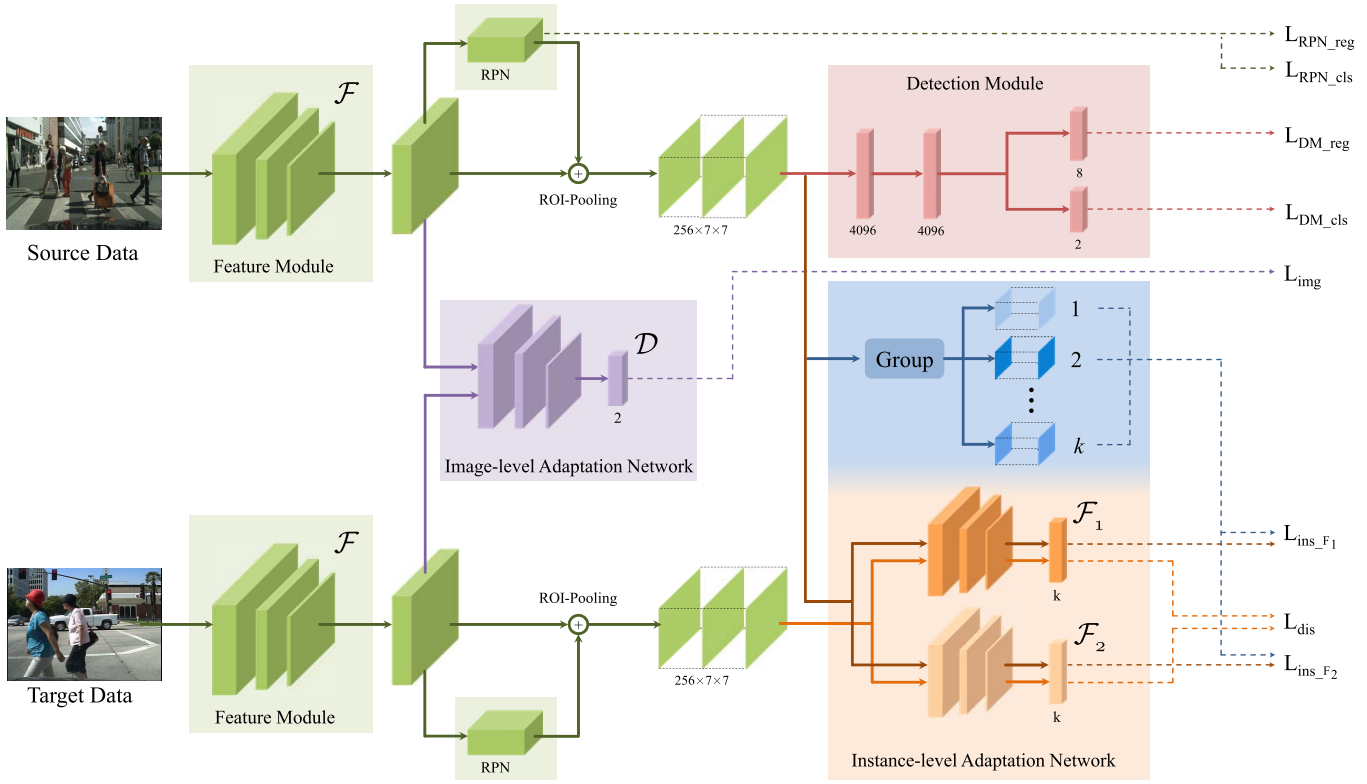


Fig. 4. The detail of the proposed model, which consists of three main parts: a Base Detector, an Image-Level Adaptation Network, and an Instance-Level Adaptation Network. For the Base Detector, the Feature module $\mathcal{F}$ is used to extract the feature map of a given image, and the RPN is applied to generate several candidate proposals, while the Detection module targets to predict the location of pedestrians along with their corresponding labels. The generated feature map is then fed into the Image-Level Adaptation Network, including a domain classifier $\mathcal{D}$ for domain alignment. The Instance-Level Adaptation Network firstly groups the candidate proposals with a Group module, and then applies the instance-level domain adaptation on the corresponding groups in the source and target domains by two classifiers $\mathcal{F}_1$ and $\mathcal{F}_2$ of Alignment module. Notably, the parameters are shared between the modules with the same color.

source and target domains, can be reduced by the *image-level domain adaptation*. Besides the image-level domain adaptation, the *instance-level domain adaptation* is applied to reduce the local visual bias, which can further align the instance description between source and target domains.

By considering the above mentioned issues, we propose a novel Selective Alignment Network to reduce the global visual bias and local visual bias for cross-domain pedestrian detection. As shown in Figure 4, the Selective Alignment Network consists of three main components: a Base Detector, an Image-Level Adaptation Network, and an Instance-Level Adaptation Network. The Base Detector, which is composed of a Feature module, an RPN module and a Detection module, is treated as a pedestrian detector and inferred with the source data. The goal of Base Detector is to generate the descriptions for images and candidate proposals, which are further applied to reduce the image bias and instance bias. After that, the Image-Level Adaptation Network and Instance-Level Adaptation Network are proposed to address image-level and instance-level domain shifts, respectively. In the following, we will give a detailed description of each component.

B. Base Detector

Similar to existing pedestrian detection approaches [50]–[52], the goal of Base Detector is to infer a robust pedestrian detector with the annotated source data $\mathcal{D}_s$. Since we need to reduce the instance bias for cross-domain pedestrian detection, the Base Detector can be implemented with the existing proposal-based pedestrian detection module. For simplification, we implement the Base Detector based on Faster R-CNN [53], which consists of a Feature module, an RPN module and a Detection module. Given an input image x, the Feature module $\mathcal{F}$ targets to extract the global image description $\mathbf{f}^g \in \mathbb{R}^{C\times W\times H}$, where C, W, and H denote the corresponding channel, width, and height. Once obtaining the global image description $\mathbf{f}^g$, the RPN module is applied to generate a set of candidate proposals $\mathbf{P} = \{p_1, p_2, \ldots\ldots, p_n\}$ based on the given anchors. After that, the feature for each proposal p_i is obtained by cropping the proposal p_i from the global description $\mathbf{f}^g$, and a following ROI-pooling operation is used to refine the proposal feature. Finally, the proposal feature is fed into the Detection module, which consists of several fully-connected layers, for pedestrian prediction and bounding box regression. Similar to the Faster R-CNN, the Base Detector can be optimized by minimizing the following loss function:

$$\mathbf{L}_{det} = L_{RPN\_cls} + L_{RPN\_reg} + L_{DM\_cls} + L_{DM\_reg}, \quad (1)$$

where L_{RPN_cls} and L_{DM_cls} are the cross-entropy loss for classification in the RPN module and Detection module,

respectively. The L_{RPN_reg} and L_{DM_reg} are the $L1$ loss for bounding box regression. More detailed description of each loss please refer to the Faster R-CNN [53].

C. Image-Level Adaptation Network

Recently, a lot of methods have been proposed for pedestrian detection to boost the performance. However, the pedestrian detector inferred from the annotated source dataset always obtains a worse performance on the target domain because source and target domains have an obvious difference. As shown in Figure 1, the detector inferred on the Caltech dataset [13] generates many false detection results on the CityPersons [14]. To improve the generalization of the pedestrian detector, the critical component is how to reduce the domain bias between the source and target domains. Inspired by the unsupervised representation learning [54], [55] and domain adaptation [56]–[59], considering the unlabeled images during training is an effective way to reduce the gap between source and target domains. Also, by considering the domain-adversarial learning [15], [20], [60], iteratively optimizing the feature generator and discriminator can align the features of source and target domains. Therefore, we propose an Image-Level Adaptation Network to minimize the approximated domain discrepancy of image descriptions. Inspired by the existing methods, the framework of domain adaptation consists of two components: feature generator and feature discriminator. In this work, we treat the Feature module $\mathcal{F}$ as the feature generator, and the feature discriminator is the Image-Level Adaptation Network $\mathcal{D}$.

Given the source images $\mathcal{X}_s$ and target images $\mathcal{X}_t$, the core of the domain classifier $\mathcal{D}$ is to be able to distinguish the source and target samples by fixing the feature generator $\mathcal{F}$. Once setting the labels for source samples and target samples as 1 and 0, we constrain the domain classifier to output a higher prediction score for the source samples and a smaller prediction score for the target samples. For the image x, the prediction score can be denoted as $\log \mathcal{D}(\mathcal{F}(x))$. Therefore, the above goal can be formulated as follows:

$$\mathcal{L}_{img} = \mathbb{E}_{x\in\mathcal{X}_s}[\log \mathcal{D}(\mathcal{F}(x))] + \mathbb{E}_{x\in\mathcal{X}_t}[\log(1 - \mathcal{D}(\mathcal{F}(x)))], \quad (2)$$

where $\mathbb{E}$ denotes the expectation.

Firstly, we minimize the objective function to optimize the discriminator $\mathcal{D}$,

$$\min_{\mathcal{D}} \mathcal{L}_{img}. \quad (3)$$

After optimizing Eq. (3), the domain classifier $\mathcal{D}$ is discriminative for the source and target samples. The goal of Image-Level Adaptation Network is to make the source features and target features have a similar distribution. That is, the source features and target features are indistinguishable. Therefore, the feature generator $\mathcal{F}$ should make the generated features indistinguishable for the inferred domain classifier $\mathcal{D}$, *e.g.,* a lower prediction score for the source samples, and a higher prediction score for the target samples. The goal can be achieved by maximizing Eq. (2) with Eq. (4),

$$\max_{\mathcal{F}} \mathcal{L}_{img}. \quad (4)$$

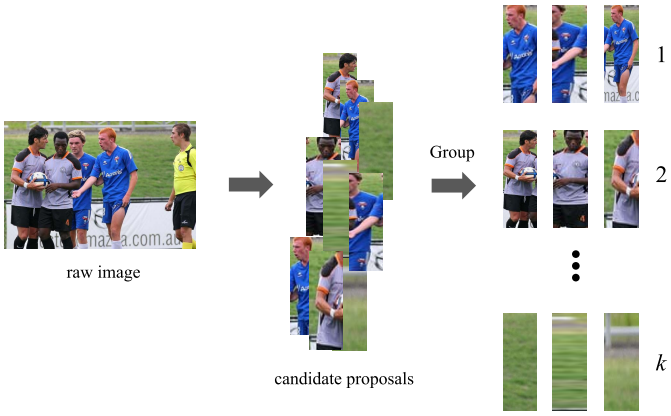


Fig. 5. The candidate proposals, generated from the raw image, are clustered into several groups with pseudo labels generated from 1 to k according to the visual features.

By iteratively optimizing Eq. (3) and Eq. (4), the feature generator $\mathcal{F}$ can align the global-level source features and target features.

D. Instance-Level Adaptation Network

Aligning image-level feature distribution alleviates the shift of style and illumination for the image, while detection is more challenging to be concerned about local instances. Therefore, how to reduce domain bias at the instance level is another crucial problem for cross-domain pedestrian detection. Inspired by the object domain adaptation [15]–[20], most existing methods [21]–[25] reduce the instance-level domain bias by aligning whole source candidate instances and whole target candidate instances, *e.g.,* Cai *et al.* [24] embed the Faster R-CNN into Mean Teacher framework with object relations to tackle the instance-level domain adaptation. The disadvantage of these methods is that they all ignore the inter-class difference of the instances belonging to the same image or domain. Since the complexity of instance appearance, the candidate instances belonging to the same image have a large inter-instance difference. To reduce the effect caused by inter-instance difference, we apply a Group module to cluster the candidate instances into several groups. Since the candidate instances belonging to each group have a similar visual description, using the Alignment module to align the instances belonging to the same group not only aligns features between two domains, but also maintains the visual discriminative, as shown in Figure 2.

Given a source image x_s along with its candidate instances $\mathbf{P}_s = \{p_1^s, p_2^s, \ldots\ldots, p_n^s\}$, where $p_i^s \in \mathbb{R}^{1\times 4}$ denotes the coordinate for each instance, and n is the number of candidate instances, the Group module aims to cluster those n instances into k groups based on the visual representation. For the instance p_i^s, its feature $\mathbf{f}_i^s \in \mathbb{R}^{C'\times W'\times H'}$ is obtained with the ROI-pooling operation. A global average pooling operation is further applied to transform $\mathbf{f}_i^s$ into a C'-dimensional feature vector. Therefore, the final description for each instance is denoted as $\mathbf{f}_i^s \in \mathbb{R}^{1\times C'}$. Once obtaining the whole instance descriptions $\mathbf{F}_s = \{\mathbf{f}_1^s, \mathbf{f}_2^s, \ldots\ldots, \mathbf{f}_n^s\}$, we apply the K-means [61] to cluster the description into k groups, as shown in Figure 5. Based on the generated groups, we can

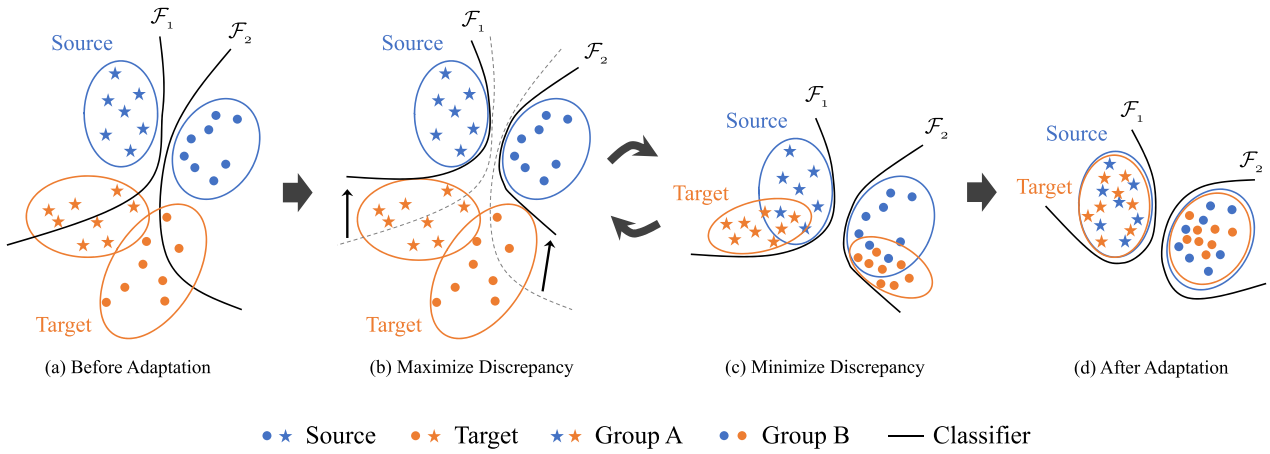


Fig. 6. The training procedure of the Alignment module. Discrepancy refers to the disagreement between the predictions of two classifiers. (a) Two classifiers ($\mathcal{F}_1$ and $\mathcal{F}_2$) can classify instances of source domain correctly with the supervision of pseudo labels. (b) We maximize the discrepancy to move the classification boundaries from dotted line to solid line. (c) By minimizing the discrepancy, the feature of instances in the target domain is getting similar generally with that of the source domain. (d) All candidate instances on the corresponding groups in the source and target domains are aligned.

define the pseudo label set $\mathbf{L} = \{l_1, l_2, \ldots\ldots, l_n\}(l_i \in [1, k])$ for all instances, where the pseudo label is the index of group. Since the group is clustered based on the visual description, the instances with the same pseudo label have similar visual features with a small visual bias.

After generating the pseudo labels for all source candidate instances, the instance-level domain adaptation can be formulated as follows. As introduced above, the source candidate instances belong to the k classes with the ground-truth labels $\mathbf{L}$. Similar to the source instances, assuming that the target candidate instances come from the same k classes, but the class labels cannot be accessed. Therefore, we propose an Alignment module to align the source and target candidate instances based on the assumption that those instances belong to the k classes. Specially, any target instance can be classified into one group of the source domain. If the target instance is very dissimilar to the source instances, it will be considered as a noise sample. Based on the assumption that most target instances might be correctly classified, those noise samples have little impact on transfer learning. As shown in Figure 4, the Alignment module consists of two classifiers $\mathcal{F}_1$ and $\mathcal{F}_2$. Since two classifiers are independent, the discrepancy between the output of two classifiers can be regarded as a supervisor to constrain the feature generator $\mathcal{F}$ to generate the alignment description. Given a target candidate instance p_i^t, the discrepancy is defined as:

$$\mathcal{L}_{dis} = \frac{1}{k}\sum_{j=1}^{k} |\mathcal{F}_1(\mathbf{f}_i^t)^j - \mathcal{F}_2(\mathbf{f}_i^t)^j|, \tag{5}$$

where $\mathbf{f}_i^t$ is the feature of p_i^t obtained from the feature generator $\mathcal{F}$.

Fixing the feature generator and maximizing the discrepancy $\mathcal{L}_{dis}$ can adjust and enlarge the classification boundaries of two classifiers, while fixing two classifiers and minimizing the discrepancy $\mathcal{L}_{dis}$ can constrain the feature generator to generate alignment features that cannot be distinguished by two classifiers. Since the classifiers are trained with annotated source images, minimizing the discrepancy can align the generated source and target features. An intuitive illustration is shown in Figure 6.

For the source candidate instances, the classifiers $\mathcal{F}_1$ and $\mathcal{F}_2$ are trained with the help of generated pseudo label $\mathbf{L}$ by minimizing the loss $\mathcal{L}_{ins}$,

$$\mathcal{L}_{ins} = \mathcal{L}_{ins\_F_1} + \mathcal{L}_{ins\_F_2}, \tag{6}$$

where $\mathcal{L}_{ins\_F_1}$ and $\mathcal{L}_{ins\_F_2}$ are the classification loss for classifiers $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively.

$$\mathcal{L}_{ins\_F_1} = -\sum_{j=1}^{k} \mathbb{1}(l = j) log \mathcal{F}_1(\mathbf{f}^s)^j, \tag{7}$$

$$\mathcal{L}_{ins\_F_2} = -\sum_{j=1}^{k} \mathbb{1}(l = j) log \mathcal{F}_2(\mathbf{f}^s)^j, \tag{8}$$

where $\mathbb{1}(\cdot)$ is a sign function.

Based on the discrepancy $\mathcal{L}_{dis}$ and the classification loss $\mathcal{L}_{ins}$, we can optimize the feature generator $\mathcal{F}$ and classifiers. Firstly, we optimize the classification loss $\mathcal{L}_{ins}$ on the source images $\mathcal{X}_s$ along with their pseudo label $\mathbf{L}$ to constrain the two classifiers that can distinguish the source images. Since two classifiers are distinct, they can easily generate a similar classification boundary for target images. To increase the gap between the boundaries of the two classifiers, we next maximize the discrepancy $\mathcal{L}_{dis}$ by fixing the feature generator. Thus, we combine the discrepancy $\mathcal{L}_{dis}$ and the classification loss $\mathcal{L}_{ins}$ to adjust the classifiers,

$$\mathcal{L}_2 = \mathcal{L}_{ins} - \mathcal{L}_{dis}. \tag{9}$$

Since two classifiers are discriminative enough, minimizing the discrepancy $\mathcal{L}_{dis}$ can constrain the two classifiers to have a similar prediction for the same unlabeled image. Once treating the trained classifiers as the feature distribution for the source domain, we finally minimize the discrepancy $\mathcal{L}_{dis}$ on target images to constrain the feature generator $\mathcal{F}$ to generate the aligned feature, as shown in Figure 6 (c).

In summary, the instance-level adaptation network aims to generate the discriminative and alignment features for source and target domains by iteratively updating the feature generator $\mathcal{F}$ and classifiers $\mathcal{F}_1$, $\mathcal{F}_2$. By fixing the feature generator, maximizing loss $\mathcal{L}_{dis}$ can adjust the discrepancy of classifier boundary to detect the target samples excluded by the support of the source. The inferred classifier can correctly classify the source samples and exclude a lot of target samples based on the fixed feature generator. Once obtaining the new two classifiers, we minimize the discrepancy $\mathcal{L}_{dis}$ by adjusting the feature generator to make the feature of each target instance move toward to the nearest source clusters which have a similar visual description with target instances. With the newly discriminative classifiers and feature generator, the unlabeled target samples can be classified into k obvious clusters according to the source clusters.

After iteratively performing the above process, the target candidate instances can be aligned with the one group of the source candidate instances. The detailed training procedure will be described in Section III-E.

E. Optimization

The full objective is to update the three components, *i.e.,* Base Detection, Image-Level Adaptation Network, and Instance-Level Adaptation Network. By taking the above mentioned constraints into consideration, the final optimization is described as follows.

Step A. Based on the annotated source images, we firstly train the Base Detector and two classifiers in Instance-Level Adaptation Network to detect and classify the source samples correctly. This step is crucial for cross-domain pedestrian detection because it can provide a robust pedestrian detector and classifiers. Therefore, the objective is a combination of Eq. (1) and Eq. (6),

$$\min_{\mathcal{F},\mathcal{D},\mathcal{F}_1,\mathcal{F}_2,} \mathcal{L}_{det} + \mathcal{L}_{ins}. \tag{10}$$

Step B. By fixing the Feature module $\mathcal{F}$, we next train the Domain Classifier $\mathcal{D}$, and the classifiers $\mathcal{F}_1$ and $\mathcal{F}_2$ to increase the discrepancy between two classifiers. The above two goals are achieved by minimizing Eq. (2) and Eq. (9),

$$\min_{\mathcal{D}} \mathcal{L}_{img}, \tag{11}$$

$$\min_{\mathcal{F}_1,\mathcal{F}_2} \mathcal{L}_{ins} - \mathcal{L}_{dis}. \tag{12}$$

Step C. By fixing the Image-Level and Instance-Level Adaptation Networks, we train the Feature module $\mathcal{F}$ to make the Domain Classifier $\mathcal{D}$ not to distinguish the unlabeled target samples, and thus minimize the discrepancy between two classifiers to generate alignment features. The above goals are optimized by maximizing Eq. (2) and minimizing Eq. (5),

$$\max_{\mathcal{F}} \mathbf{L}_{img}, \tag{13}$$

$$\min_{\mathcal{F}} \mathbf{L}_{dis}. \tag{14}$$

F. Training and Testing

During the training phase, the Base Detector, Image-Level Adaptation Network, and Instance-Level Adaptation Network are all trained with the source images along with theirs ground-truth annotations, and the target images. Notably, both Adaptation Networks can guide to optimize the Feature module $\mathcal{F}$ via back-propagation to reduce the bias between source and target domains. During the test phase, only the Base Detector is applied for pedestrian detection benefiting from the learned domain-invariant features.

IV. Experiments

A. Datasets

To demonstrate the effectiveness of the proposed Selective Alignment Network, we evaluate the proposed method on three datasets: Caltech [13], CityPersons [14], and COCOPersons [26]. The detailed description of each dataset is shown as follows.

Caltech. The Caltech dataset [13] consists of about 250, 000 frames from approximately 10 hours of 640×480 30Hz video taken in an urban environment. Each pedestrian is labeled with a bounding box of the full extent of the entire pedestrian. Furthermore, the visible parts for the occluded pedestrians are provided for three classes, *i.e.,* 'person', 'people', and 'person?'. The individual pedestrians are labeled 'Person'. Large groups of pedestrians or impossible to label individuals are delineated using a single bounding box and labeled as 'People'. In addition, the label 'Person?' is assigned when clear identification of a pedestrian is ambiguous or easily mistaken. Among the three types of images, we only use 'person' for training. The 42,782 images coming from the set00-set05 are used for training, and the 4,024 images coming from set06-set10 are used for evaluation.

CityPersons. The CityPersons dataset [14] is built upon the Cityscapes [62]. Similar to the Caltech dataset [13], the labeled pedestrians are classified into four classes, *i.e.,* 'pedestrian', 'rider', 'sitting person', and 'other person', and each pedestrian is labeled with a bounding box of both full bodies and visible parts. In this work, we only use the 'pedestrian'. There are approximately 5,000 2048×1024 images in total, including 2,975 training images, 500 images for validation, and 1,525 images for test.

COCOPersons. The COCOPersons dataset [26] is a subset of MSCOCO detection dataset by only considering the 'person' and ignoring the other 79 classes. Finally, there are 64,115 training images and 2,639 test images. Different from the annotations for Caltech and CityPersons, the persons in COCOPersons are annotated as the visible body.

For cross-domain pedestrian detection, one of them is used as the source dataset, and the rest can be treated as the target domain. Some examples of three datasets are illustrated in Figure 7, which demonstrate the visual bias among different datasets.

B. Evaluation Settings and Metrics

1) Settings: With the provided three datasets, we establish four types of settings for cross-domain adaptation:

(a) Caltech

(b) CityPersons

(c) COCOPersons

Fig. 7. Samples for Caltech(a), CityPersons(b), and COCOPersons(c).

1) *Caltech-to-CityPersons*: treating the Caltech as the source dataset and CityPersons as the target dataset; (2) *CityPersons-to-Caltech*: treating the CityPersons as the source dataset and Caltech as the target dataset; (3) *COCOPersons-to-Caltech*: treating the COCOPersons as the source dataset and Caltech as the target dataset; (4) *COCOPersons-to-CityPersons*: treating the COCOPersons as the source dataset and CityPersons as the target dataset.

2) Metrics: Similar to the official metric for Caltech and CityPersons datasets, the log-average miss rate, which is computed in the False Positive Per Image (FPPI) range of $[10^{-2}, 10^{0}]$ (denoted as MR^{-2}), is used to measure the detection performance.

C. Implementation Details

1) Use of Training Data: Since the images belonging to the COCOPersons have different resolutions, we thus resize them into 640×480. By performing the transfer between source and target datasets, all target images are resized to the same resolution as that in the source dataset. During training the Base Detector, we treat the candidate proposal whose IoU is larger than 0.5 with any ground-truth as the positive sample, and whose maximum IoU with all ground-truth ranges from 0.1 to 0.3 as the negative sample. As the number of negative samples is far more than that of the positive samples, we use all positive samples and randomly select several negative samples to make up 256 samples in total per image for training.

2) Initialization and Setting for Training: We implement the proposed method with PyTorch on NVIDIA TITAN Xp GPU. The Stochastic Gradient Descent (SGD) is applied to optimize the network, which adopts random initialization without any pre-trained model. The gamma, decay step and decay ratio are set as 0.9, 5 and 0.1. The learning rate is set as 0.001, and adjust every 5 epochs. We use the batch size $=$ 1, *i.e.,* one source image and one target image. The proposed model is stopped after about 146,000 iterations for Caltech-to-CityPersons transfer, 35,000 iterations for CityPersons-to-Caltech transfer, 422,000 iterations for COCOPersons-to-CityPersons transfer, and 328,000 iterations for COCOPersons-to-Caltech transfer.

TABLE I

EFFECT OF BASE DETECTOR ON CITYPERSONS-TO-CALTECH ADAPTATION

Methods	Base Detector	$MR^{-2}(\%)$
ours	FRCNN (ResNet-50)	24.41
ours	FRCNN (VGG-16)	**14.27**

D. Baselines

In this work, we analyze and compare the proposed method with the following baselines to verify its effectiveness.

1) **FRCNN:** The original Faster R-CNN model is trained with source data and directly evaluated on target dataset.
2) **FRCNN_IM:** The model consists of the Base Detector and the Image-Level Adaptation.
3) **FRCNN_BIN:** The model consists of the Base Detector, the proposed Image-Level Adaptation module, and the Instance-Level Adaptation module introduced in DAFR [21].
4) **SAN:** The model consists of the Base Detector, the Image-Level Adaptation Network and the Instance-Level Adaptation Network as described in Sec. III-C and Sec. III-D, respectively.

E. Ablation Studies

The contributions of the proposed Selective Alignment Network (SAN) include introducing two constraints to reduce the domain bias between the source and target domains: 1) using *image-level adaptation network* to reduce the global visual bias; 2) using *instance-level adaptation network* to reduce the local visual bias. We thus analyze the effectiveness of each constraint.

1) Effect of Base Detector: We firstly analyze the effect of the base detector, and summarize the related results in Table I. As shown in Table I, the proposed method with base detector FRCNN(VGG-16) obtains better performance than that with the detector FRCNN(ResNet-50), *e.g.,* reducing the Miss Rate from 24.41% to 14.27% on City→Cal. The reason is that the down-sampling rate of ResNet-50 at convolution layers is too large for the network to detect and localize small pedestrians.

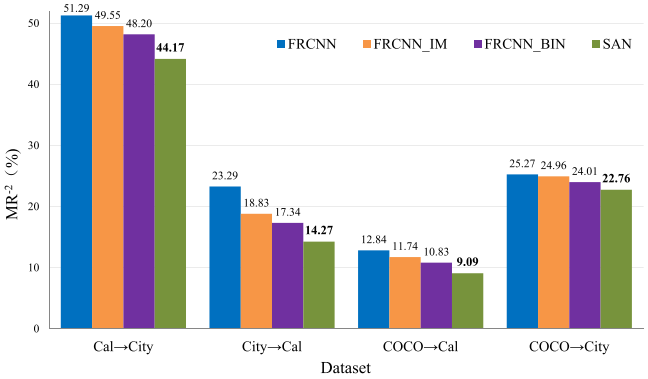


Fig. 8. Effect of Image-Level Adaptation Network and Instance-Level Adaptation Network on four types of settings for cross-domain adaptation scenarios. $MR^{-2}(\%)$ is used to measure the detection performance based on the new annotations [6]. Cal→City, City→Cal, COCO→Cal and COCO→City stand for Caltech-to-CityPersons, CityPersons-to-Caltech, COCOPersons-to-Caltech and COCOPersons-to-CityPersons, respectively.

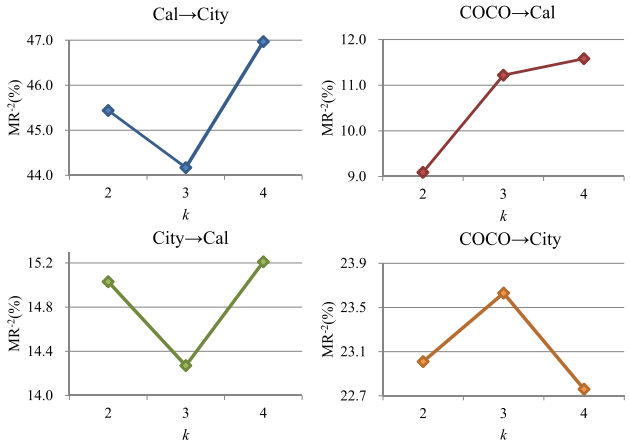


Fig. 9. Effect of Group Number on four types of settings for cross-domain adaptation scenarios.

2) Effect of Image-Level Adaptation Network: We secondly analyze the effect of image-level adaptation network, and summarize the related results in Figure 8. As shown in Figure 8, the FRCNN_IM with image-level adaptation network obtains the detection performance improvement upon the FRCNN on four types of settings, *e.g.,* reducing the Miss Rate from 51.29%, 23.29%, 12.84%, and 25.27% to 49.55%, 18.83%, 11.74% and 24.96% for Caltech-to-CityPersons, CityPersons-to-Caltech, COCOPersons-to-Caltech and COCOPersons-to-CityPersons, respectively. The better performance demonstrates that using image-level adaptation network can alleviate the shift of style and illumination of the image, and reduce the global visual bias between the source and target domains in cross-domain pedestrian detection.

3) Effect of Instance-Level Adaptation Network: We also evaluate the effect of instance-level adaptation network, and summarize the detailed results in Figure 8. As shown in Figure 8, the FRCNN_BIN, which considers the additional basic instance-level alignment, obtains a lower Miss Rate than the FRCNN_IM. The lower Miss Rate proves that using the basic instance-level alignment is critical to cross-domain pedestrian detection. Note that the basic instance-level alignment applies the whole candidate instances to align two domains. Different from considering the whole candidate instances, we propose a group-based instance alignment to reduce the instance-inter difference. From Figure 8, we can observe that the SAN, which is proposed with the group-based instance alignment, is superior to the FRCNN_BIN on all four settings, *e.g.,* reducing the Miss Rate from 48.20%, 17.34%, 10.83% and 24.01% to 44.17%, 14.27%, 9.09% and 22.76% for Caltech-to-CityPersons, CityPersons-to-Caltech, COCOPersons-to-Caltech and COCOPersons-to-CityPersons, respectively. The better performance demonstrates the advantage of considering the inter-instance difference for instance-level domain adaptation compared with aligning all candidate instances.

4) Effect of Group Number: As discussed above, the group-based instance alignment is an effective way to align the candidate instances between source and target domains. Since the group number k plays a critical role in group-based instance alignment, we thus analyze the effect of group number k. The analysis is performed in terms of Miss Rate (MR^{-2}), and results are shown in Figure 9. As shown in Figure 9, setting k as 3 yields the best performance for Caltech-to-CityPersons and CityPersons-to-Caltech, while $k = 2$ achieves the lowest miss rate for COCOPersons-to-Caltech. For the COCOPersons-to-CityPersons, setting k as 4 obtains the best performance. The above comparison shows that different domain adaptation settings should use different group numbers because each dataset contains different pedestrian images. As shown in Figure 7, the visual description of pedestrians in the COCOPersons dataset is more diverse, which leads to more groups that should be considered. We further give some visualization results for different group number. As shown in Figure 10, using the Group module can divide the complicated instances well, and overcome the misalignment for source and target images.

5) Effect of Instance Number: As described in Section IV-C, we select 256 candidate instances per image for the source domain. Furthermore, we perform several experiments to explore the effect of the number of instance for source and target domains, and choose the best hyper-parameter instance number N_{ins} used in this work. The evaluation is performed based on CityPersons-to-Caltech with three kinds of Group Number, *i.e.,* $k = 2, 3$, and 4, and the results are summarized in Table II. As shown in Table II, setting the instance number as 256 obtains the best performance compared with other two settings. The reason is that considering too more or too fewer instance number N_{ins} for target domain would lead to imbalanced samples in each clustered group compared with the source domain, which can seriously affect the learning of domain-invariant features.

6) Effect of Iteration Number: Since we need to iteratively update the Feature module $\mathcal{F}$ and Domain Adaptation models, *e.g.,* Domain Classifier $\mathcal{D}$ in Image-Level Adaptation Network,

Fig. 10. Illustration of the grouping results for different group number. (a) raw image, (b) representative instances after grouping with group number $k = 2$, (b) representative instances after grouping with group number $k = 3$.

TABLE II
EFFECT OF INSTANCE NUMBER ON CITYPERSONS-TO-CALTECH ADAPTATION UNDER THREE KINDS OF GROUP NUMBER ($k = 2, 3, 4$)

Methods	$MR^{-2}(\%)$
ours(k=2, N_{ins}=128)	16.85
ours(k=2, N_{ins}=256)	**15.85**
ours(k=2, N_{ins}=512)	16.02
ours(k=3, N_{ins}=128)	16.41
ours(k=3, N_{ins}=256)	**15.40**
ours(k=3, N_{ins}=512)	16.01
ours(k=4, N_{ins}=128)	16.99
ours(k=4, N_{ins}=256)	**15.94**
ours(k=4, N_{ins}=512)	16.04

TABLE III
EFFECT OF ITERATION NUMBER ON CITYPERSONS-TO-CALTECH ADAPTATION UNDER THREE KINDS OF GROUP NUMBER ($k = 2, 3, 4$)

Methods	$MR^{-2}(\%)$
ours(k=2, N_{gen}=1, N_{dis}=1)	15.85
ours(k=2, N_{gen}=2, N_{dis}=1)	15.25
ours(k=2, N_{gen}=3, N_{dis}=1)	**15.03**
ours(k=3, N_{gen}=1, N_{dis}=1)	15.40
ours(k=3, N_{gen}=2, N_{dis}=1)	15.06
ours(k=3, N_{gen}=3, N_{dis}=1)	**14.27**
ours(k=4, N_{gen}=1, N_{dis}=1)	15.94
ours(k=4, N_{gen}=2, N_{dis}=1)	15.38
ours(k=4, N_{gen}=3, N_{dis}=1)	**15.21**

and two classifiers $\mathcal{F}_1$ and $\mathcal{F}_2$ in Instance-Level Adaptation Network. Inspired by the generative adversarial learning, we update the Feature module N_{gen} times and the Domain Adaptation models N_{dis} times during a training session. For example, $N_{gen} = 2$ and $N_{dis} = 1$ stand for that the Feature module $\mathcal{F}$ iterates two epochs, and the $\mathcal{D}$, $\mathcal{F}_1$ and $\mathcal{F}_2$ iterates one epoch. We thus analyze the effect of N_{gen} and N_{dis}, and summarize the results in Table III. As shown in Table III, increasing the number of N_{gen} can reduce the Miss Rate, *e.g.,* setting $N_{gen} = 3$ and $N_{dis} = 1$ obtains the lowest Miss Rate among all settings. The reason is that the feature generation plays a critical role in domain adaptation, and using a higher N_{gen} can constrain the Feature module to generate more discriminative alignment descriptions.

F. Comparisons With State-of-the-Arts

In this section, we compare the proposed method with existing methods on three datasets, *i.e.,* Caltech [13], CityPersons [14] and COCOPersons [26].

Table IV summarizes the comparisons between the SAN and several existing methods, which are trained with source data only and directly evaluated on the target domain, *i.e.,* FRCNN [53], Cas-RCNN [63], ALFNet [33], and FPN [64]. From Table IV, we can see that the proposed method achieves the best performance compared with the existing methods, *e.g.,*

TABLE IV

COMPARISONS OF STATE-OF-THE-ART METHODS ON FOUR TYPES OF SETTINGS FOR CROSS-DOMAIN ADAPTATION SCENARIOS. $MR^{-2}(\%)$ IS USED TO MEASURE THE DETECTION PERFORMANCE BASED ON THE NEW ANNOTATIONS [6]. CAL→CITY, CITY→CAL, COCO→CAL AND COCO→CITY STAND FOR CALTECH-TO-CITYPERSONS, CITYPERSONS-TO-CALTECH, COCOPERSONS-TO-CALTECH AND COCOPERSONS-TO-CITYPERSONS, RESPECTIVELY. BOLD NUMBER INDICATES THE BEST RESULT

Methods	Cal→City	City→Cal	COCO→Cal	COCO→City
FRCNN [53]	51.29	23.29	12.84	25.27
Cas-RCNN [63]	47.36	16.73	12.12	24.87
ALFNet [33]	45.52	18.30	20.03	40.69
FPN [64]	45.92	14.69	10.90	38.62
DAFR [21]	56.68	18.42	17.31	50.18
SCDA [22]	51.25	28.93	12.18	29.33
SAN	**44.17**	**14.27**	**9.09**	**22.76**

reducing the Miss Rate of 1.3%, 0.4%, 1.8%, and 2.1% on four adaptation scenarios, respectively. As shown in Table IV, the ALFNet [33] obtains a worse performance for the domain adaptation setting when treats the COCOPersons as the source dataset, *e.g.,* the Miss Rates are 20.03% and 40.69% for COCO→Cal and COCO→City, respectively. The reason is that the ALFNet is proposed for the traditional pedestrian detection, which may not be suitable for the COCOPersons.[1]

As shown in Table IV, we also observe that the existing object or pedestrian detectors all obtain a high Miss Rate on Cal→City. The reason is that they are simply designed for the pedestrian detection whose training and test data have a similar distribution. Without the learning ability for domain-invariant features, those methods are not suitable for recognizing pedestrians on unseen domains, thus resulting in worse detection performance. To make up for the above deficiency, our proposed SAN can address the domain shift well.

We also compare the proposed method with several cross-domain detection methods, *i.e.,* DAFR [21], and SCDA [22], and summarized the related results in Table IV. As shown in Table IV, the proposed method obtains better performance than DAFR and SCDA, *e.g.,* reducing the Miss Rate of 12.51%, 4.15%, 8.22%, 27.42% and 7.08%, 14.66%, 3.09%, 6.57% on Cal→City, City→Cal, COCO→Cal and COCO→City, respectively. The reason is that DAFR and SCDA simply consider the whole candidate proposals and location information for domain alignment. Therefore, we can conclude that using the group-based instance alignment is an effective way to align the source and target domains.

V. CONCLUSION

Cross-domain pedestrian detection aims to recognize pedestrians belonging to the unseen target domain with the help of labeled source domain. In this work, we propose a novel Selective Alignment Network (SAN) for cross-domain pedestrian detection, which consists of three sub-networks: a Base Detector, an Image-Level Adaptation Network and an Instance-Level Adaptation Network. The goal of Base Detector is to generate a robust detector and descriptor for source images, which are further applied to reduce the image bias and instance bias between two domains. After that, the Image-Level Adaptation Network and Instance-Level Adaptation Network are proposed to address image-level and instance-level domain shifts, respectively. Extensive experiments with various settings among three challenging datasets demonstrate the effectiveness of the proposed method.

In this work, we have proved that aligning the local instance by considering the inter-instance difference is a more effective way for cross-domain pedestrian detection than merely considering the whole instances. For the proposed Selective Alignment Network, we apply the K-means to cluster the candidate instances, which is an efficient way. Therefore, how to make a more detailed division becomes a significant problem for cross-domain detection. We will conduct this research in the future.

[1]We reimplement the ALFNet with the provided code on the COCOPersons dataset.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 26th Annu. Conf. Neural Inf. Process. Systems.*, 2012, pp. 1106–1114.

[2] R. Benenson, M. Omran, J. H. Hosang, and B. Schiele, "Ten years of pedestrian detection, what have we learned?" in *Proc. Eur. Conf. Comput. Vis., Workshops*, 2014, pp. 613–627.

[3] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "Towards reaching human performance in pedestrian detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 973–986, Apr. 2018.

[4] J. Hosang, M. Omran, R. Benenson, and B. Schiele, "Taking a deeper look at pedestrians," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4073–4082.

[5] Y. Tian, P. Luo, X. Wang, and X. Tang, "Pedestrian detection aided by deep learning semantic tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5079–5087.

[6] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1259–1267.

[7] C. Zhou and J. Yuan, "Multi-label learning of part detectors for heavily occluded pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3506–3515.

[8] W. Ouyang, X. Zeng, and X. Wang, "Partial occlusion handling in pedestrian detection with a deep model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 11, pp. 2123–2137, Nov. 2016.

[9] Q. Hu, P. Wang, C. Shen, A. van den Hengel, and F. Porikli, "Pushing the limits of deep CNNs for pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1358–1368, Jun. 2018.

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

[11] 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.

[12] H. Yao and C. Xu, "Joint person objectness and repulsion for person search," *IEEE Trans. Image Process.*, vol. 30, pp. 685–696, 2021.

[13] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.

[14] 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.

[15] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 1180–1189.

[16] J. Hoffman *et al.*, "Cycada: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 80, 2018, pp. 1994–2003.

[17] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[18] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3723–3732.

[19] C. Tao, F. Lv, L. Duan, and M. Wu, "Minimax entropy network: Learning category-invariant features for domain adaptation," *CoRR*, vol. abs/1904.09601, 2019.

[20] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2962–2971.

[21] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3339–3348.

[22] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, "Adapting object detectors via selective cross-domain alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 687–696.

[23] T. Wang, X. Zhang, L. Yuan, and J. Feng, "Few-shot adaptive faster R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7173–7182.

[24] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao, "Exploring object relation in mean teacher for cross-domain detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11457–11466.

[25] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa, "Cross-domain weakly-supervised object detection through progressive domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5001–5009.

[26] T. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV), Part V*, in Lecture Notes in Computer Science, vol. 8693. Zürich, Switzerland: Springer, Sep. 2014, pp. 740–755.

[27] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[28] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 424–432.

[29] 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.

[30] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[31] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–11.

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

[33] 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.*, 2018, pp. 643–659.

[34] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for pedestrian detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 745–761.

[35] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 985–996, Apr. 2018.

[36] 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.

[37] C. Zhou, M. Yang, and J. Yuan, "Discriminative feature transformation for occluded pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9556–9565.

[38] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5187–5196.

[39] Y. Jiao, H. Yao, and C. Xu, "PEN: Pose-embedding network for pedestrian detection," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jun. 5, 2020, doi: 10.1109/TCSVT.2020.3000223.

[40] B. Geng, D. Tao, and C. Xu, "DAML: Domain adaptation metric learning," *IEEE Trans. Image Process.*, vol. 20, no. 10, pp. 2980–2989, Oct. 2011.

[41] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2066–2073.

[42] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, 2015, pp. 97–105.

[43] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[44] H. Xia and Z. Ding, "Hgnet: Hybrid generative network for zero-shot domain adaptation," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV), Part XXVII*, in Lecture Notes in Computer Science, vol. 12372. Glasgow, U.K.: Springer, Aug. 2020, pp. 55–70.

[45] Y. Zuo, H. Yao, and C. Xu, "Category-level adversarial self-ensembling for domain adaptation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[46] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. Macready, "A robust learning approach to domain adaptive object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 480–490.

[47] Z. He and L. Zhang, "Multi-adversarial faster-RCNN for unrestricted object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6667–6676.

[48] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6956–6965.

[49] Z. He and L. Zhang, "Domain adaptive object detection via asymmetric tri-way faster-RCNN," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV), Part XXIV*, in Lecture Notes in Computer Science, vol. 12369. Glasgow, U.K.: Springer, Aug. 2020, pp. 309–324.

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

[51] 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.

[52] G. Brazil, X. Yin, and X. Liu, "Illuminating pedestrians via simultaneous detection and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4960–4969.

[53] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[54] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, 2016.

[55] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, 2018.

[56] Y. Chen, S. Song, S. Li, and C. Wu, "A graph embedding framework for maximum mean discrepancy-based domain adaptation algorithms," *IEEE Trans. Image Process.*, vol. 29, pp. 199–213, 2020.

[57] Y. Liu, W. Tu, B. Du, L. Zhang, and D. Tao, "Homologous component analysis for domain adaptation," *IEEE Trans. Image Process.*, vol. 29, pp. 1074–1089, 2020.

[58] A. Chadha and Y. Andreopoulos, "Improved techniques for adversarial discriminative domain adaptation," *IEEE Trans. Image Process.*, vol. 29, pp. 2622–2637, 2020.

[59] B. Gholami, P. Sahu, O. Rudovic, K. Bousmalis, and V. Pavlovic, "Unsupervised multi-target domain adaptation: An information theoretic approach," *IEEE Trans. Image Process.*, vol. 29, pp. 3993–4002, 2020.

[60] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4068–4076.

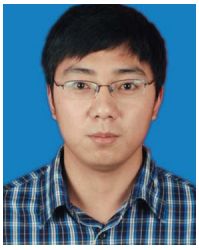
[61] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable K-means++," *Proc. VLDB Endow.*, vol. 5, no. 7, pp. 622–633, 2012.

[62] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

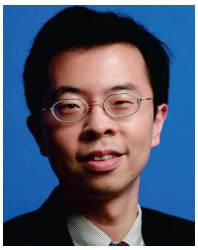
[63] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[64] 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.

Yifan Jiao received the B.S. degree from Jiangsu University, Zhenjiang, China, in 2015, and the M.S. degree from the Jiangsu University of Science and Technology, Zhenjiang, China, in 2018. He is currently pursuing the Ph.D. degree in technology of computer application with the Hefei University of Technology, Hefei, China. His research interests include computer vision and multimedia.



Hantao Yao (Member, IEEE) received the B.S. degree from Xidian University, Xi'an, China, in 2012, and the Ph.D. degree from the Institute of Computing Technology, University of Chinese Academy of Sciences, in 2018. After graduation, he worked as a postdoctoral, from 2018 to 2020, at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He is currently an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He was a recipient of National Postdoctoral Programme for Innovative Talents. His current research interests are zero-shot learning, person tracking and detection, and person re-identification.



Changsheng Xu (Fellow, IEEE) is currently a Professor with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences and Executive Director of the China-Singapore Institute of Digital Media. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition, and computer vision. He has hold 30 granted/pending patents and published over 200 refereed research papers in these areas. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing, Communications and Applications and ACM/Springer Multimedia Systems Journal*. He received the Best Associate Editor Award of *ACM Transactions on Multimedia Computing, Communications and Applications*, in 2012 and the Best Editorial Member Award of *ACM/Springer Multimedia Systems Journal* in 2008. He served as Program Chair of ACM Multimedia 2009. He has served as associate editor, guest editor, general chair, program chair, area/track chair, special session organizer, session chair and TPC member for over 20 IEEE and ACM prestigious multimedia journals, conferences and workshops. He is IAPR Fellow and ACM Distinguished Scientist.