Unsupervised Pedestrian Detection with Progressive Domain Transfer

Zhiri Tang[a], Jian Zhong[a], Qianfen Jiao[a], Si Wu[b], Hau-San Wong[a,*]

[a]*Department of Computer Science, City University of Hong Kong, Hong Kong*
[b]*Department of Computer Science, South China University of Technology, China*

Abstract

The performance of current pedestrian detector relies heavily on the number of annotations. Although sufficient labels can help to improve the performance, the generalization ability of well-trained models is usually not good enough in a new domain. Hence, how to bridge the gap between source and target domains is still a challenge for pedestrian detection. In this work, an unsupervised pedestrian detection framework with progressive domain transfer is proposed. First, a scene translation module is applied to generate an intermediate domain, in which the images and labels of source domain match with similar distributions of target domain. To avoid instability in the translation process, a blind image quality assessment module is applied to weight images in the intermediate domain to avoid the influence of low-quality images on the training process. In addition, an anchor-free detector with domain confusion regularization is adopted to achieve better generalization in the target domain. Experimental results on a number of benchmark pedestrian detection datasets show that the proposed framework can effectively improve the unsupervised detection performance in the target domain, and represent a feasible way for cross-domain unsupervised detection.

Keywords: pedestrian detection, domain transfer, scene translation, image quality assessment, anchor-free detector

*Corresponding author
Email address: cshswong@cityu.edu.hk (Hau-San Wong)

1. Introduction

Pedestrian detection is one of the most significant research topics in computer vision, through which the classification and localization of pedestrians can be achieved. Although pedestrian detection has been applied in many applications such as video surveillance [1] and automatic driving [2], the performance of detection relies heavily on the number of annotations. When faced with scenes in a target domain without annotations, the performance of a well-trained detector in the original source domain often has a significant decrease of performance due to differences between the domains such as viewpoints, resolutions, pedestrian density, and so on. Although providing some labeled training samples in the target domain can effectively improve the performance and help to reduce the distribution gap between domains, significant manpower and resources are still required for collecting annotations.

Unsupervised domain adaptation (UDA) aims to bridge the gap between different domains, which usually needs annotations from the source domain and some specific distributions of source and target domains [3]. Recent works on UDA are implemented for pattern recognition tasks [4] while few works on other complex tasks such as object detection [5] and semantic segmentation [6] are considered.

Based on the above, we formulate the solution of the domain shift problem as a progressive domain transfer task, and an overview of the proposed framework is shown in Fig. 1. First, an intermediate domain between source and target domains is generated by an unsupervised image-to-image translation module. This module can be regarded as a bridge between source and target domains so that the original task can be converted into a cross-domain task from the intermediate domain to the target domain. Since some images in the intermediate domain are of low quality, a scoring process for all the images in the intermediate domain is applied via a blind image quality assessment module, where the resulting scores are used as weights for the training data. In other words, the contributions of low-quality images can be reduced to avoid negative

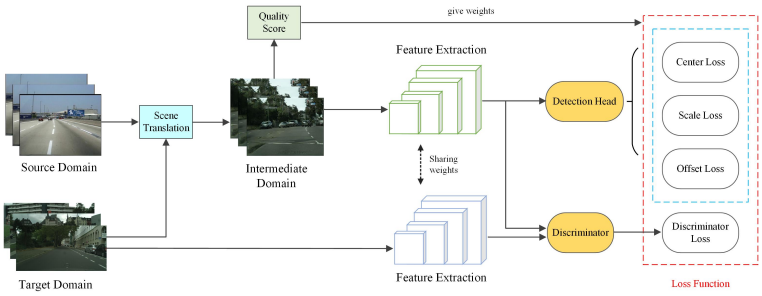

Figure 1: An overview of the proposed unsupervised pedestrian detection framework with progressive domain transfer. First, an intermediate domain is generated by a scene translation module, where an image quality assessment module is used for scoring all the images in the intermediate domain. Then, an anchor-free detector with domain confusion regularization is used as the detection backbone of this work, where the entire loss function consists of detection and discriminator losses with weights from image quality assessment.

influence on cross-domain performance via the scoring process. In addition, an anchor-free detector is applied as the backbone, which uses a set of detection heads to obtain the center and scale of pedestrians. This backbone is different from a traditional anchor-based detector which can obain the detection results without a set of pre-defined anchors. Furthermore, a discriminator is used for domain confusion regularization which reduces the gap between intermediate and target domains. This process can also be regarded as a domain adaptation process in the feature space. To evaluate the performance of the proposed framework, experiments are conducted based on three common pedestrian detection datasets including Caltech [7], CityPersons [8], and KITTI [9]. Experimental results show that the proposed framework can effectively improve the unsupervised detection performance under the cross-domain setting. The main contributions can be summarized as follows: (1) An intermediate domain generated by a scene translation module is used as a bridge between source and target domains, where all the images in the intermediate domain are scored by a blind image quality assessment module to mitigate influence from low-quality images; (2) An anchor-free detector with domain confusion regularization based on a discriminator is designed for domain adaptation; (3) A number of benchmark

pedestrian detection datasets are used to evaluate the performance, which shows the effectiveness of the proposed unsupervised pedestrian detection framework.

2. Related works

2.1. Object detection

With the development of deep learning, computer vision becomes one of the important research topics in recent years. As one of the high-level tasks in computer vision, object detection includes two subtasks, which are classification and localization of target objects. Traditional object detection relies on predefined anchors with different sizes and scales to obtain the height and width of target instances. The mainstream frameworks can be divided into two types: two-stage and one-stage anchor-based detectors. Faster R-CNN [10], which is representative works of two-stage detectors, generate proposals for targets to obtain classification results. While achieving high accuracy, the processing speed of two-stage detection frameworks is also relatively slow. One-stage detectors, such as YOLO [11] and SSD [12] skip the step of proposal generation to help improve the processing speed.

Apart from the above anchor-based detectors, a number of new anchor-free detectors emerge in recent years. CornerNet [13] applies a keypoint-based method to obtain the top-left and bottom-right corner of a target from the attention maps directly. FSAF [14] presents a set of anchor-free branches to replace traditional anchor-based branches based on two subnetworks for classification and regression. Similarly, FCOS [15] proposes a detection head to perform classification and regression tasks based on multi-level inference. Furthermore, CSP [16] also presents a detection head to obtain the center and scale of targets from concatenated feature maps. Although the recent works on anchor-free detection achieve state-of-the-art accuracy and speed, the performance relies heavily on the number of labeled data. When the distribution of unlabeled images in the target domain is different form that of source domain, the performance of the above well-trained models will significantly decrease. Hence, the design of a

backbone with better generalization ability on the target domain represents a major challenge in cross-domain object detection.

2.2. Unsupervised pedestrian detection

Unsupervised pedestrian detection is an important task in many applications. The scene-specific detector presented by Zeng et al. [17] aims to learn multi-scale features to improve the feature representation and generalization of the entire framework. Another approach to design scene-specific detector is proposed by Wang et al. [18], which includes a four-step transfer learning process. However, the above two approaches are complex. To address this issue, a deep domain adaptation framework using Maximum Mean Discrepancy (MMD) loss [19] is presented for unsupervised pedestrian detection [20]. Furthermore, a self-paced learning method with a progressive latent model [21] is also a feasible approach for detection in real-world datasets.

In recent years, works on unsupervised pedestrian detection can be divided into two mainstream categories, which are retraining with noisy labels and adding training data using a generation network. RoyChowdhury et al. [22] presented an automatic adaptation detection framework to the target domain, which uses a well-trained detector from the source domain to obtain noisy labels in the target domain. Then noisy labels with high confidence scores are applied for retraining. Besides, using a generation network to synthsize more training data is also a common way to improve the unsupervised detection performance. Hsu et al. [23] applied GAN to generate a new domain between source and target domains, which includes source data and labels with distribution of the target domain. Another work [24] applied GAN to generate new pedestrians from source data for training, which can provide richer and more complex features in the training data. Although these works improve performance to different extents, it is still a challenging task to design more effective and efficient detection frameworks.

2.3. Domain adaptation

Domain adaptation is proposed to bridge the gap between source and target domains with few or even no labels in the latter. Recently, a set of works are presented based on adversarial learning methods. These works include DANN (domain adversarial neural networks) [25], DSN (domain separation networks), [26] and domain adaptative faster r-cnn [27] which use adversarial learning as their backbone for domain adaptation. While the above works carry out feature-level adaptation, there are also other works which focus on image-level adaptation. For image classification and semantic segmentation, CycleGAN [28] and related GAN [4] framework are applied for image-to-image translation. The synthesized images are used for training, which incorporate labels from source domain and a similar data distribution of the target domain. Most works about object detection and pedestrian detection are under a weakly supervised [29] or semi-supervised [30] setting, and they have comparable performance with fully supervised frameworks. Only a few works focus on cross-domain unsupervised detection [20, 23].

3. Proposed approach

3.1. Scene translation with scoring

3.1.1. Learning an intermediate domain

To bridge the gap between source and target domains, an intermediate domain, in which the distributions of source and target images can be matched, is applied to convert the original problem into a cross-domain task between intermediate and target domains. Inspired by [31], an unsupervised image-to-image translation network is used to generate an intermediate domain with a shared-latent space assumption. The translation network assumes that a pair of images from source and target domains can be represented in the same latent space, based on which it can generate corresponding images from different domains. An overview of generating an intermediate domain with samples from CityPersons (Domain A) and Caltech (Domain B) are shown in Fig. 2. In this

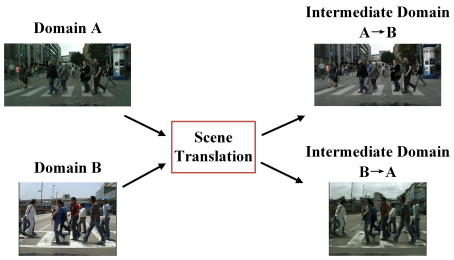

Figure 2: An overview of learning an intermediate domain. Domain A and B represent the CityPersons and Caltech datasets, respectively. Take intermediate domain A→B as example, it has similar scene distributions with domain B. More importantly, the localization, size, and shape of pedestrians in images from the intermediate domain A→B have no changes compared with images in domain A, which provides strong justifications to use the original labels of domain A for training.

example, two intermediate domains are generated based on the original images in domain A and B, where the intermediate domain A→B can be regarded as a new domain with similar distributions as domain B and the labels of domain A. In other words, the original problem of bridging the gap between A and B can be decomposed into two sub-problems: from domain A to intermediate domain and from intermediate domain to domain B. The first sub-problem, which is from domain A to intermediate domain, is implemented by the scene translation network.

3.1.2. Scoring generated images

Although the above scene translation process can reduce the discrepancy between the distributions of source and target images, the instability of the translation network may lead to degradation of detection performance. Some samples in the intermediate domains are shown in Fig. 3, in which some distortions can be observed in the sky, and some fake pedestrians can be seen on the road in the target-like source domain in Fig. 3 (a). These low-quality images will affect the training process and the detection performance. In Fig. 3 (b), the target-like images are with high quality, which can be regarded as a reliable intermediate domain for training.

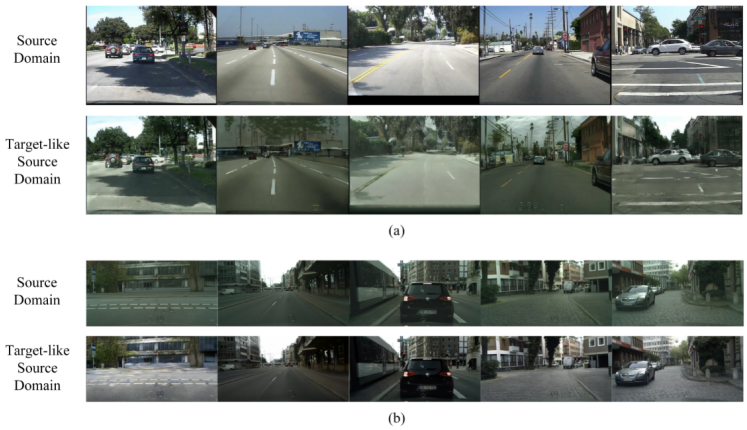

Figure 3: Some samples in the intermediate domains: (a) dataset A to dataset B and (b) dataset B to dataset A. In (a), the sky of the second and fourth images in the target-like source domain are distorted compared with the original images. Furthermore, a new pedestrian has appeared in the left road of the third image, and an original pedestrian has disappeared in the left wall of the fifth image in the generated domain. In (b), all the five sample images have no obvious distortion.

To solve this issue, a blind image quality assessment (IQA) module is applied to give scores for images in the generated domain. The module adopts a no-reference IQA method, which means it does not need prior knowledge of the distortions in natural images or a high-quality image for reference. Different from other no-reference IQA methods which require knowledge on anticipated distortions in the training dataset and pre-scoring by human users, the IQA module measures only deviations from statistical information in images under assessment without any training process or obtaining distributions of unknown images. Overall, it is suitable for giving scores for the generated images in the intermediate domain and reducing negative influence from low-quality images.

$$D(\mu_1, \mu_2, \sigma_1, \sigma_2) = ((\mu_1 - \mu_2)^T (\frac{\sigma_1 + \sigma_2}{2})^{-1} (\mu_1 - \mu_2))^{\frac{1}{2}} \quad (1)$$

where μ_1 and μ_2 are the means of multivariate Gaussian distributions of natural image and distorted image, respectively. Furthermore, σ_1 and σ_2 are the standard deviations of the two distributions.

To weight the translated images for training, a min-max normalization is performed as follows:

$$w(I) = 1 - \frac{D - D_{min}}{D_{max} - D_{min}} \tag{2}$$

where $w(I)$ is the weight value for a translated image. The above equation indicates that higher weights are assigned to high-quality images. Through the scoring process, weights for images in the intermediate domain can help to avoid the instability of scene translation and achieve consistent cross-domain detection performance.

3.2. Anchor-free detector with domain confusion

3.2.1. Detection network

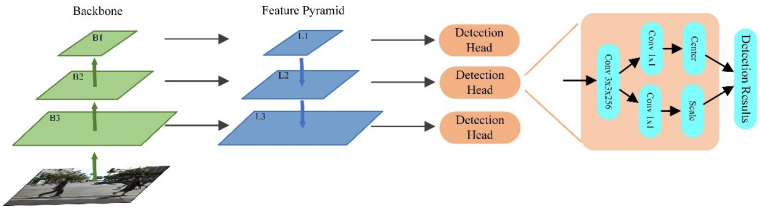

Figure 4: The detection network module.

An anchor-free detector with multi-scale inference is applied as the backbone detection network in this work, which is shown in Fig. 4. A set of feature maps with different sizes are presented as $\{L_1, L_2, L_3\}$, which are given from the backbones $\{B_1, B_2, B_3\}$ by a 1×1 convolution kernel following a top-down structure. The strides of L_1, L_2, and L_3 are 8, 16, and 32, respectively.

For each layer in the feature pyramid, a detection head is applied to obtain the center and scale of objects, which has a 3×3 convolutional layer and two 1×1 convolutional layers to extract the center heatmap and scale respectively. For center regression, a cross-entropy loss is assigned. To model the ambiguity of the center point, a Gaussian mask $G(.)$ is assigned for location. Hence an

overall mask is presented as

$$Mask_{ij} = \max_{n=1,...N} G(i, j; x_n, y_n, \delta_{w_n}, \delta_{h_n}) \tag{3}$$

with the Gaussian function as follows:

$$G(i, j; x_n, y_n, \delta_{w_n}, \delta_{h_n}) = e^{-(\frac{(i-x)^2}{2\delta_w^2} + \frac{(j-y)^2}{2\delta_h^2})} \tag{4}$$

In the above equation, N is the total number of pedestrians in the image, (x_n, y_n, w_n, h_n) is the location of the center, and δ_{w_n} and δ_{h_n} are proportional to the height and width of a single pedestrian, respectively.

To address the class imbalance problem between positive and negative samples, a focal loss [32] is applied to formulate the classification loss function of the center as follows:

$$L_c = -\frac{1}{N} \sum_{i=1}^{W/r} \sum_{j=1}^{H/r} \alpha_{ij} (1 - p_{ij})^{\gamma} \log(p_{ij}) \tag{5}$$

where

$$p_{ij} = \begin{cases} p_{ij} & y_{ij} = 1 \\ 1 - p_{ij} & otherwise \end{cases} \tag{6}$$

$$\alpha_{ij} = \begin{cases} 1 & y_{ij} = 1 \\ (1 - M_{ij})^{\beta} & otherwise \end{cases} \tag{7}$$

The probability of the target pedestrian in location (i, j) is $p_{ij} \in [0, 1]$ and the label is $y_{ij} \in \{0, 1\}$. Further, the hyper-parameters α and β are set to 2 and 4 based on experimental observations, which is also consistent with [32] and [33].

The scale regression is given by a smooth L1 loss, which is always applied in object detection tasks for robust and fast convergence:

$$L_s = \frac{1}{N} \sum_{n=1}^{N} SmoothL1(s_n, gt_n) \tag{8}$$

where s_n and gt_n are the network output and actual label of pedestrians, respectively.

Similarly, an offset loss, which is proposed for improving the accuracy of bounding boxes, is also formulated via smooth L1 loss as follows:

$$L_o = \frac{1}{N}\sum_{n=1}^{N} SmoothL1(x_n/r, x_n^*/r) + \frac{1}{N}\sum_{n=1}^{N} SmoothL1(y_n/r, y_n^*/r) \quad (9)$$

where x_n/r and y_n/r are predictions from the network, and x_n^*/r and y_n^*/r are the ground truth.

Hence, the entire loss function of detector is given by:

$$L_{det} = \lambda_c L_c + \lambda_s L_s + \lambda_o L_o \quad (10)$$

where λ_c, λ_s, and λ_o are weights for the center, scale, and offset terms, respectively.

During the inference process, each level of the feature pyramid gives prediction via the corresponding detection head. To avoid ambiguous detection results, a hard threshold for the scale at specific levels is applied [15]. In most cases, the overlap of different predictions is from different levels of the feature map. If the overlapped bounding boxes correspond to the same object, the prediction result with the smallest scale is preserved and the other boxes are deleted.

3.2.2. Domain confusion

We propose to apply domain confusion regularization to the detector, which uses a discriminator after the feature pyramid to judge whether the input is from source or target domain. An overview of the domain confusion module is shown as Fig. 5.

Let P_t be the estimated probability of the input image belonging to target domain, and d is a label where $d = 1$ indicates that the image is from the target domain, while $d = 0$ indicates that it originates from the source domain. The loss function of the discriminator is formulated as follows:

$$L_{dis}(F(I)) = -\sum d \log P_t + (1-d)\log(1-P_t) \quad (11)$$

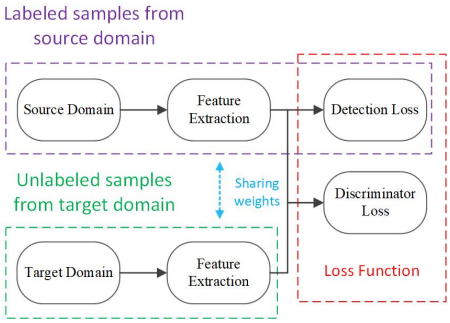

Figure 5: An overview of the domain confusion module.

Since the training goal is to confuse the model such that it cannot judge which domain the input images come from, we need to maximize the discriminator loss during the training process. The detection and discriminator frameworks are parameterized by θ_{det} and θ_{dis}, respectively. Combining with the detection loss function, the complete loss function is expressed as:

$$\min_{\theta_{det}} \ \lambda_{det} L_{det} + \lambda_{dis} L_{dis} \tag{12}$$

$$\max_{\theta_{dis}} \ \lambda_{dis} L_{dis} \tag{13}$$

where λ_{det} and λ_{dis} are hyper-parameters to trade off the impacts of the various losses.

3.3. Complete framework

Following the above steps, an intermediate domain via scene translation and scoring based on image quality assessment are presented first. Then an anchor-free detector with multi-scale inference with domain confusion is introduced. Based on the scoring process and the loss function of detector with domain confusion, the final training loss can be formulated as:

$$\min_{\theta_{det}} \ w(I) \ (\lambda_{det} L_{det} + \lambda_{dis} L_{dis}) \tag{14}$$

$$\max_{\theta_{dis}} \; w(I)\; \lambda_{dis} L_{dis} \tag{15}$$

where $w(I)$ is the weight value from eq. (2) for training images. From the loss function, a weighted training process is developed, which can effectively improve the training stability when using generated images from the intermediate domain and help to avoid the negative influence from low-quality training images.

4. Experiments

4.1. Experimental setup

4.1.1. Dataset

To evaluate performance of the proposed method, three common pedestrian detection datasets including Caltech [7], CityPersons [8], and KITTI [9] are selected for test and comparisons. The Caltech dataset contains 42,782 training images and 4,024 test images taken from vehicle-mounted camera in an urban environment, in which a total of 2,300 pedestrians are annotated. Furthermore, a new sanitized set of annotations [34] are adopted for evaluation.

KITTI dataset contains 7,481 high-resolution images in the training set. However, because there are no public annotations in the test set of KITTI, the training set is randomly divided into two subsets for training and testing, where the ratio is 2:1 following [8].

The training set of CityPersons is obtained from 18 cities and contains 2,975 images and 19,654 persons with annotations. In the validation set, 500 images and 3,938 persons are obtained from 3 different cities. Compared with Caltech and KITTI, CityPersons is a relatively difficult dataset with a much more complex environment and more pedestrians.

4.1.2. Implementation details

ResNet-50 [35] is chosen as the backbone for the proposed detection network and the architecture of the discriminator follows that of [23] including three

convolution layers with 64 channels and one additional layer for binary classification. The parameters in the detection network, including λ_c, λ_s, and λ_o are 0.01, 1, and 0.1, respectively, following the settings in [16]. Furthermore, the weight of the discriminator λ_{dis} is set to 0.1. The intermediate domain is generated by UNIT [31] and the image quality assessment process follows the NIQE framework [36]. The batch size is one during the entire training process and the proposed framework is implemented using one RTX 2080 Ti GPU.

4.2. Experimental results and analysis

4.2.1. Performance of scene translation with scoring

Some samples from the source domain and intermediate domain are shown in Fig. 6. The first column of Fig. 6 (a) includes three images from the Caltech dataset, and the second and third columns are corresponding images in two different intermediate domains, which are CityPersons-like and KITTI-like (Cal2City and Cal2KIT), respectively. From the samples, the images in the intermediate domain have similar distributions with the target domain including season, weather, and brightness, while the locations and sizes of pedestrians in these generated images are preserved. Another intermediate domain is shown as the second column in Fig. 6 (b), which is a Caltech-like CityPersons dataset (City2Cal).

As can be seen from the figure, using the intermediate domain as training set can lead to better cross-domain performance. To show the merits of training on the intermediate domain, t-SNE [37] is applied to show the feature space of the source domain, intermediate domain, and target domain. Three hundred random images from Caltech, CityPersons, and Cal2City datasets are used as examples and the visualization results are shown in Fig. 7. It can be observed that the samples from Cal2City and CityPersons are difficult to classify compared with the Caltech dataset. Hence, using Cal2City as training set can help to achieve better generalization performance in the CityPersons dataset.

To avoid the influence of low-quality images in the intermediate domain, a scoring process based on blind image quality assessment framework is proposed

Figure 6: Some samples from source domain and intermediate domain. In (a), images in the first column are from the original Caltech dataset. The second and third columns are from Cal2City and Cal2KIT, respectively. In (b), images in the first column are from the original CityPersons dataset and the second column is from City2Cal.

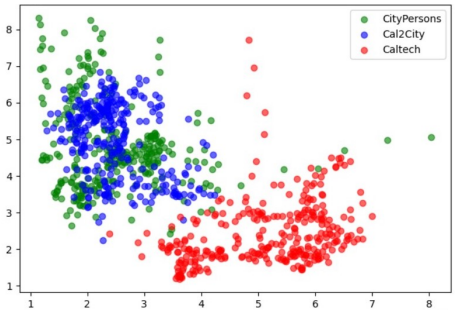

Figure 7: Visualization results of Caltech, Cal2City, and CityPersons datasets via t-SNE, which shows 300 random samples selected from the three different domains.

and three samples in Cal2City dataset are shown in Fig. 8. Fig. 8 (a) is with relatively high image quality with $D_1 = 13.170$ and $w_1 = 0.932$ based on Eqs. (1) and (2). In Fig. 8 (b), there are many blurred regions in the sky but no obvious distortions on the road. Hence the quality of this image is medium with $D_2 = 14.087$ and $w_2 = 0.770$. In Fig. 8 (c), there are many new blurred pedestrians or pedestrian-like objects on the road. These new objects, especially those that are very similar to pedestrians, will negatively affect the training process and make the detector confused about whether these objects are pedestrians or not. As a result, the score of the image is low with $D_3 = 17.403$ and $w_3 = 0.250$. It can be seen that the proposed method can effectively assign weights for the generated images and reduce the contributions of low-quality images to the training process, which leads to better performance and avoids the influence from instability of the scene translation framework.

(a) (b) (c)

Figure 8: Three samples in Cal2City dataset, whose weights assigned by the proposed method are (a) 0.932, (b) 0.770, and (c) 0.250, respectively.

Based on the above, the weight distributions of the three intermediate domains are shown in Fig. 9. From the distributions, weights of most images fall within $[0.5, 0.9]$ and only a few are with very low weights (< 0.5) or high weights (> 0.9).

Some examples of detection results before and after scene translation with scoring (STrans-scoring) from Caltech to CityPersons are shown in Fig. 10. Fig. 10 (a) and (b) represent detection results after ST-scoring. Fig. 10 (c) and (d) represent results after using scene translation without the scoring process (STrans). Fig. 10 (e) and (f) are detection results from the CSP detector [16],

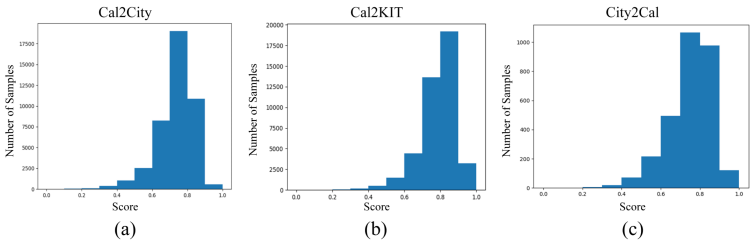

Figure 9: The distributions of training weights of generated images in (a) Cal2City, (b) Cal2KIT, and (c) City2Cal.

which are used as comparisons. Green bounding boxes are correct detection results given by the corresponding models. Yellow and red bounding boxes represent incorrect detection results and missed pedestrians, respectively. As can be observed, the adoption of STrans leads to more green boxes and fewer yellow and red ones compared with detection results given by the CSP detector. Furthermore, the proposed STrans-scoring can help to achieve better detection performance.

Besides the above detection results in some images from CityPersons, evaluations following standard Caltech metric [7], which is log-average Miss Rate over False Positive Per Image (FPPI) in $[10^{-2}, 10^{0}]$ (denoted as MR^{-2}) are shown in Table 1. In the table, "CSP [16]" are trained with source data only and the results of "Oracle" are obtained under a fully supervised setting. Ablation studies under different settings are also conducted on Cal2City and City2Cal to verify the effectiveness of the proposed methods. "STrans" presents the performance using scene translation to generate an intermediate domain as training data. Compared with the original CSP detector, our method leads to improved performance by approximately 5-6% in Cal2City and 2-3% in City2Cal. Furthermore, "STrans-scoring" denotes the detection results of the entire scene translation with scoring process, which shows better performance compared with others. For Cal2City, STrans-scoring can lead to improvement by about 10-11% compared with CSP and about 4-5% compared with STrans. For City2Cal, STrans-scoring can improve performance by about 4-5% compared with the original

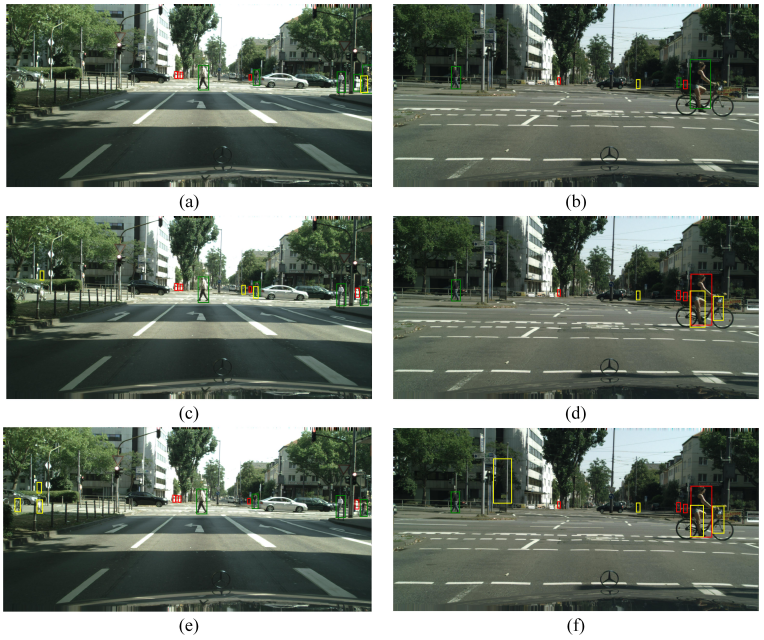

Figure 10: Comparison of detection results before and after applying the proposed STrans-scoring from Caltech to CityPersons. (a) and (b) are detection results after STrans-scoring. (c) and (d) are detection results after using STrans only. (e) and (f) are detection results from the CSP detector [16]. (Green bounding boxes represent correct detection results. Yellow and red boxes represent incorrect detection results and missed pedestrians, respectively.)

results given by CSP.

Caltech → CityPersons		CityPersons → Caltech	
Method	$MR^{-2}(\%)$	Method	$MR^{-2}(\%)$
CSP [16]	53.98	CSP [16]	21.64
STrans	48.41	STrans	18.85
STrans-scoring	**43.57**	STrans-scoring	**17.63**
Oracle	11.0	Oracle	4.54

Table 1: Unsupervised detection results using STrans-scoring. The best experimental results are shown in bold.

4.2.2. Performance of anchor-free detector with domain confusion

Some examples of detection results before and after applying the proposed anchor-free detector with domain confusion (AFDete-dc) from Caltech to CityPersons are shown in Fig. 11. Fig. 11 (e) and (f) are detection results from the CSP detector [16] for comparison. Fig. 11 (c) and (d) represent results after using the proposed anchor-free detection framework without domain confusion (AFDete). Compared with the CSP detector, the proposed framework can detect more correct pedestrians. Fig. 11 (a) and (b) present detection results based on AFDete-dc, including the proposed detection framework and domain confusion. It can be observed that our framework leads to a higher detection accuracy as indicated by the increased number of green bounding boxes.

We present values of $MR^{-2}(\%)$ of three different cross-domain settings with ablation studies in Table 2. "AFDete" and "AFDete-dc" denote the detection results of the proposed detection framework without and with domain confusion, respectively. For Cal2City, AFDete and AFDete-dc improve performance by about 3-4% and 7-8%, respectively, compared with CSP. For City2Cal, the proposed AFDete-dc also improves the performance by 3-4%.

Figure 11: Comparison of detection results before and after applying the proposed AFDete-dc from Caltech to CityPersons. (a) and (b) are detection results after AFDete-dc. (c) and (d) are detection results after AFDete. (e) and (f) are detection results from the CSP detector [16].

Caltech → CityPersons		CityPersons → Caltech	
Method	$MR^{-2}(\%)$	Method	$MR^{-2}(\%)$
CSP [16]	53.98	CSP [16]	21.64
AFDete	50.35	AFDete	19.34
AFDete-dc	**46.73**	AFDete-dc	**18.19**
Oracle	11.0	Oracle	4.54

Table 2: Unsupervised detection results using AFDete-dc. The best experimental results are shown in bold.

4.2.3. Comparison with existing methods

For Cal2City, two images from CityPersons dataset are selected to show the detection performance before and after applying the entire proposed framework with progressive domain transfer (PDT), which is shown in Fig. 12. In Fig. 12 (a) and (c), which is a sample image with few pedestrians, the proposed PDT can detect all the five pedestrians while the original CSP detector misses two of them under the cross-domain setting. In Fig. 12 (b) and (d), which is a sample with many pedestrians, the proposed PDT detects three more pedestrians and reduces the number of incorrect detection by three. For City2Cal, the performance of the proposed framework with ablation studies is shown in Fig. 13. It can be observed from the figure that our framework outperforms the CSP detector under the unsupervised setting by 7-8%.

Figure 12: Comparison of detection results before and after applying the proposed PDT from Caltech to CityPersons. (a) and (b) are detection results after STrans-scoring and ACDete-dc. (c) and (d) are detection results from CSP detector [16].

The unsupervised detection performance on City2Cal are shown in Table 3, where "PDT" denotes the results from the proposed framework. We observe that PDT leads to improved performance by approximately 7-8% in City2Cal compared with the original CSP [16] detection framework. In addition, we have included other state-of-the-art works for comparison. ACF [38] is a classic

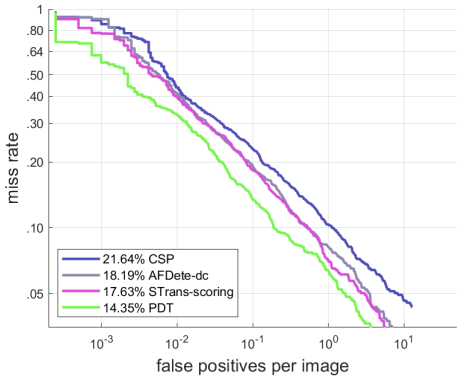

Figure 13: Comparison of detection results before and after applying PDT from CityPersons to Caltech.

detector which uses extrapolation from nearby scales to approximate features in complex images and achieve fast and accurate object detection. Adapted FasterRCNN [8] is proposed to achieve better detection accuracy and improved generalization ability across a number of different pedestrian detection datasets. ALFNet [39], which includes an Asymptotic Localization Fitting (ALF) module to present anchor boxes based on a one-stage detection framework, achieve both the detection accuracy of two-stage detectors and the processing speed of one-stage detectors. Compared with the above detection frameworks, the proposed PDT attains better performance by 36-37% compared to ACF, 6-7% compared to Adapted FasterRCNN, and 10-11% compared to ALFNet.

In addition to the above experiments, two recent frameworks aiming for better generalization ability are also selected for comparison. PRNet [40] (Progressive Refinement Network) introduces a one-stage detection framework with three progressive phases for addressing the occluded pedestrian detection problem, and for detecting pedestrians with small sizes and incomplete structure based on modified receptive fields. Experimental results in [40] show that it can achieve state-of-the-art generalization ability under the unsupervised setting. Compared with PRNet, the unsupervised detection performance on City2Cal is further improved by our proposed framework by approximately 3-4%.

APGAN [24] (Attribute Preserving Generative Adversarial Networks) aims to improve the generalization ability of pre-trained detectors for new datasets. Based on generated new pedestrians with good visual quality and style transfer, this work can effectively improve the generalization ability of well-trained models via data augmentation of generated new pedestrians in source images. Compared with APGAN, our proposed PDT can also improve the detection performance by about 6-7%. Based on the above experimental results and comparisons with state-of-the-art works on City2Cal, the proposed PDT can effectively improve the unsupervised detection performance.

CityPersons → Caltech	
Method	MR^{-2}(%)
ACF [38]	51.28
CSP [16]	21.64
Adapted FasterRCNN [8]	21.18
ALFNet [39]	25.0
PRNet [40]	18.3
APGAN [24]	20.5
PDT	**14.35**
Oracle	4.54

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

Apart from the above experimental results on City2Cal, two other settings including Cal2City and Cal2KIT are also shown in Table 4. Although only few works including ACF [38], Adapted FasterRCNN [8], and CSP [16] present the cross-domain performance for Cal2City and Cal2KIT, our work also shows superiority based on the comparisons. Specifically, the proposed PDT can improve performance by approximately 15-16% in Cal2City and 2-3% in Cal2KIT compared with the CSP detector. Additionally, compared with Adapted FasterRCNN and ACF, our work can also improve the detection performance by 8-9%

and 34-35% on City2Cal, respectively. For Cal2KIT which only has a small gap between unsupervised detection and Oracle, our work can still achieve better performance. From the above experimental results on other datasets, we can conclude that the proposed PDT can significantly improve the unsupervised pedestrian detection performance.

Caltech → CityPersons		Caltech → KITTI	
Method	$MR^{-2}(\%)$	Method	$MR^{-2}(\%)$
ACF [38]	72.89	ACF [38]	49.99
Adapted FasterRCNN [8]	46.91	Adapted FasterRCNN [8]	10.50
CSP [16]	53.98	CSP [16]	12.81
PDT	**38.42**	PDT	**10.42**
Oracle	11.0	Oracle	7.86

Table 4: Unsupervised detection performance on Cal2City and Cal2KIT, and comparisons with state-of-the-art works. The best experimental results are shown in bold.

5. Conclusion

In this work, an unsupervised pedestrian detection framework with progressive domain transfer is proposed. Scene translation with scoring enhances the training effectiveness by utilizing images from an intermediate domain, which is generated by scene translation, with the generated images weighted by a blind image quality assessment module based on their quality. Further, a new anchor-free detector with domain confusion is applied to perform domain adaptation in the feature space. Experimental results on three benchmark pedestrian detection datasets show that the proposed framework can lead to significantly better performance compared with a number of state-of-the-art detection frameworks.

In the future, there are several directions we can focus on to improve the robustness and accuracy of cross-domain pedestrian detection: 1. Fuse the image quality assessment into the generation framework as a loss function, which can generate a target-like source domain with high image quality directly. It should

be a better way rather than only score the generated images; 2. During the scene translation process, some pedestrian instances are missing although high image quality is preserved. To solve this problem, a pedestrian instance-based scene translation framework should be a better way to maintain the quality of pedestrian instances.

References

[1] X. Wei, H. Zhang, S. Liu, Y. Lu, Pedestrian detection in underground mines via parallel feature transfer network, Pattern Recognition 103 (2020) 107195.

[2] Y. Yan, M. Xu, J. S. Smith, M. Shen, J. Xi, Multicamera pedestrian detection using logic minimization, Pattern Recognition (2020) 107703.

[3] B. Yang, P. C. Yuen, Learning adaptive geometry for unsupervised domain adaptation, Pattern Recognition 110 (2021) 107638.

[4] W. Chen, H. Hu, Generative attention adversarial classification network for unsupervised domain adaptation, Pattern Recognition 107 (2020) 107440.

[5] S. Kim, J. Choi, T. Kim, C. Kim, Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6092–6101.

[6] R. Li, W. Cao, Q. Jiao, S. Wu, H.-S. Wong, Simplified unsupervised image translation for semantic segmentation adaptation, Pattern Recognition 105 (2020) 107343.

[7] 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.

[8] 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.

[9] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The kitti dataset, The International Journal of Robotics Research 32 (11) (2013) 1231–1237.

[10] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[11] 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.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.

[13] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, Centernet: Keypoint triplets for object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6569–6578.

[14] C. Zhu, Y. He, M. Savvides, Feature selective anchor-free module for single-shot object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 840–849.

[15] 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.

[16] 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.

[17] 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.

[18] X. Wang, M. Wang, W. Li, Scene-specific pedestrian detection for static video surveillance, IEEE transactions on pattern analysis and machine intelligence 36 (2) (2013) 361–374.

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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 401 (2020) 123–132.

[25] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, The Journal of Machine Learning Research 17 (1) (2016) 2096–2030.

[26] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, D. Erhan, Domain separation networks, in: Advances in Neural Information Processing Systems, 2016, pp. 343–351.

[27] Y. Chen, W. Li, C. Sakaridis, D. Dai, L. Van Gool, Domain adaptive faster r-cnn for object detection in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3339–3348.

[28] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[29] J. Zhang, H. Su, W. Zou, X. Gong, Z. Zhang, F. Shen, Cadn: A weakly supervised learning-based category-aware object detection network for surface defect detection, Pattern Recognition 109 (2021) 107571.

[30] Y. Tang, J. Wang, B. Gao, E. Dellandréa, R. Gaizauskas, L. Chen, Large scale semi-supervised object detection using visual and semantic knowledge transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2119–2128.

[31] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: Advances in Neural Information Processing Systems, 2017, pp. 700–708.

[32] 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.

[33] H. Law, J. Deng, Cornernet: Detecting objects as paired keypoints, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 734–750.

[34] 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.

[35] 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.

[36] A. Mittal, R. Soundararajan, A. C. Bovik, Making a "completely blind" image quality analyzer, IEEE Signal processing letters 20 (3) (2012) 209–212.

[37] L. v. d. Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (Nov) (2008) 2579–2605.

[38] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (8) (2014) 1532–1545.

[39] 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.

[40] 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.