Unreliability-aware Disentangling for Cross-Domain Semi-supervised Pedestrian Detection

Wenhao Wu[1], Si Wu[2,*], and Hau-San Wong[1]

[1] Department of Computer Science, City University of Hong Kong
wenhaowu5-c@my.cityu.edu.hk, cshswong@cityu.edu.hk
[2] School of Computer Science and Engineering, South China University of Technology
cswusi@scut.edu.cn

Abstract. The rapid progress of pedestrian detection is supported by the ever-growing labeled training data and elaborate neural-network-based model. However, adequate labeled training data are not always accessible when it comes to a new scene. Semi-supervised learning is promising for the case where a small amount of manually annotated images and a large amount of unannotated images are handy. In the semi-supervised setting, data generation is a powerful technique as a type of data augmentation. Some methods conduct data generation by disentangling pedestrian instances into different codes in latent space and combining codes of different instances to reconstruct new instances. However, these methods either work in a single domain or cannot handle the case where some instances are partially represented in the images. In this work, we propose to solve code-level information transferring from reliable domains to unreliable domains by incorporating a domain classifier that competes with the disentangling module to generate domain-invariant codes. An external classifier is trained on appearance-enhanced instances and sends integrity signals to the generative module, which facilitates the generative module to recognize fully/partially represented pedestrian instances. The resulting classifier ultimately renders high-quality pseudo-annotations for the unannotated data. The pseudo-annotated data, combined with a small amount of manually annotated data, are used to achieve a detector with more generalization and accuracy. We perform extensive experiments on multiple challenging benchmarks to demonstrate the effectiveness of the proposed method.

Keywords: Pedestrian Detection · Semi-supervised Learning · Domain Adaptation.

1 Introduction

Pedestrian detection, which is a fundamental task in the computer vision community and well applied to a number of practical applications, like autonomous

[*] Corresponding author.

driving and intelligent surveillance, has experienced significant progress in recent years, especially after the emergence of the deep network. However, pedestrian detection still experiences many challenges, and the existing methods mainly put their focus on learning robust features against occlusion, variant scales and illumination change. The remarkable performances of these works are at the cost of a huge number of labeled data, which is hugely time-consuming and human-resource-consuming. When it comes to limited labeled data, these performances face serious deterioration.

Semi-supervised method, which utilizes limited labeled data and a large number of unlabeled data to achieve an improved pedestrian detector, is a promising method to solve the challenge of annotated data deficiency. The key to semi-supervised pedestrian detection is to produce trustworthy annotations for unannotated images, which is used to re-training the detector combined with annotated images. Some efforts [27, 29, 28] have been proposed to improve the performance of pedestrian detectors through applying a pre-trained detector on unannotated images to extract high-confidence bounding boxes and re-training the detector. However, these methods have their limitations on the poor discrimination of the filtering mechanism over the hard positive and negative. Another attempt [33, 3] is to synthesize pedestrian instances through generative adversarial networks (GANs) [10] to improve the generalization and differentiation of discriminative modules. However, these GAN-based methods usually need exhaustive fine-tuning and are inclined to produce unreliable pedestrian instances. Therefore, we need to propose a more controllable method to produce diverse and reliable pedestrian instances for improving the recognition ability of the discriminative module.

In this work, we focus on generating pedestrian instances with manageable latent codes. DG-Net [41] is a powerful framework for joint person disentangling and re-identification in single domains. However, the success of DG-Net is at the cost of the availability of identity information and high-quality person instance images. When given low-quality and partially-represented pedestrian instances without any identity knowledge in realistic detection scenes of different domains, DG-Net performs poorly in disentangling and reconstruction. To solve this problem, we propose to disentangle pedestrian instances into id-related appearance codes and id-unrelated structure codes through shared id-related and id-unrelated encoders, respectively. Different types of codes from different instances can be united to generate abundant unseen instances in the target domains. An additional code-level domain classifier is incorporated to compete with the id-related encoder, which accounts mainly for the diversity of generated pedestrian instances, of the generative module. An external classifier is appended to the generative module to absorb the knowledge of diverse pedestrian instances and return integrity signals of given instances to the generative module. The trained classifier is well discriminative on hard positives and negatives and is used to generate highly reliable pseudo annotations for unannotated images. Our goal is to re-train the base detector with a small amount of manually annotated and a large amount of pseudo-annotated data, and encourage the resulting detector to perform as closely as the model trained on fully-annotated data.

The main contributions of this work include: (1) We handle the case of transferring knowledge from calibrated pedestrian instances to uncalibrated pedestrian instances through adversarial training between the id-related encoder and domain classifier, and integrity signals from the external classifier; (2) The classifier trained on appearance-rich data can differentiate high-quality pseudo annotations from a bunch of pedestrian/background instances, which is used to achieve a powerful re-trained detector on several benchmark datasets.

2 Related Work

2.1 Scene-specific Pedestrian Detection

The prevalence of deep neural networks drives a great advancement in both generic object detection and pedestrian detection tasks. Particularly, Faster R-CNN [25] is a revolutionary work and achieves an incredible performance on object detection. In Faster R-CNN, features generated by Region Proposal Network (RPN) are fed into the fully-connected-based classifier. Based on Faster R-CNN, a series of works have been developed on the pedestrian detection task. Zhang et al. [38] fed features generated through the RPN module into a boosted decision forest for pedestrian/background classification. Cai et al. [2] unified detections from different detection heads implemented on feature maps of different layers to achieve scale invariance. Mao et al. [23] incorporated extra segmentation information into features to enhance the semantic information. Brazil et al. [1] further enhance features with semantic information at the image-level and instance-level to improve detection performance.

Pedestrian detection has many challenges, such as scale variance and occlusion, which are also hot spots to explore. Wang et al. [30] developed a novel repulsion loss, which improves the model's regression ability and prevents predicted boxes from inaccurate shifting. Zhou and Yuan [42] instead explored the contribution of visible parts to occluded pedestrian detection and proposed a bi-box regression model. Liu et al. [20] developed an adaptive-NMS to dynamic suppress non-negative bounding boxes based on density scores generated from the density sub-network. Chi et al. [5, 6] utilized the head information, which is hardly occluded, to detect occluded pedestrians. For scale variance, Li et al. [17] concatenated features from multiple parallel sub-networks, each of which accounts for pedestrian detection of different scale ranges. Wu et al. [31] proposed to force the features of small-scale pedestrians to mimic those of large-scale pedestrians, which therefore enhances the features of small-scale pedestrians. Kim et al. [13] proposed to memorize the features of large-scale pedestrians and recall similar features whenever meeting small-scale pedestrians.

Although plenty of works were developed to solve different challenges in the pedestrian detection task, there are limited existing methods that were originally designed to concentrate on the insufficient supervision problem. Rosenberg et al.

[26] adopted the self-training method to utilize a pre-trained detector to obtain high-confidence detected boxes from unlabeled images, which are then used to re-train the detector. Wang et al. [28] transferred a detector to a new scene by applying the target samples with generated labels. Zeng et al. [37] simultaneously achieved classification and reconstruction tasks to analyze the data distribution of the target scene. Mhalla et al. [24] adopted the sequential Monte Carlo filter to estimate and approximate the data distribution in the target scene. Wu et al. [35] utilized a FCN-based verification module to improve the quality of pseudo annotations of unlabeled images. Wu et al. [36] proposed to reduce the domain gap between data from source scenes and target scenes by adversarial training in the feature space and developed a collaborative learning mechanism to make full use of the aligned source samples.

2.2 Pedestrian Synthesis

The emergence of generative adversarial networks (GANs) advances the development of image generation including pedestrian instance generation. Wu et al. [32] adopted a cascaded model on the multi-source information, including masked images, instance-level masks and edge maps, to generate scene-specific pedestrian patches. Cheung et al. [4] estimated camera parameters and the Spawn Probability Maps for given unannotated images to determine the location of generated pedestrians in scene images, constructing annotated images with generated annotations. Wu et al. [33] improved the Triple-GAN [16] to generate pedestrian instances from noise, enhancing the diversity of training samples. Lin et al. [18] translated unreliable instances generated from a pre-trained detector to reliable instances through a well-designed generator. Zheng et al. [41] proposed to disentangle person information into id-related appearance information and in-unrelated structure information, followed by combining appearance and structure information from different instances to construct brand new instances with known identity information. Zou et al. [44] further extended this method to combine information from instances of different domains to create id-known instances of a target domain. Our approach is different from these works: We focus on the case where transferring information from id-rich well-represented instances of the source domain to id-deficient partially-represented instances of the target domain.

3 Method

3.1 Overview

Our proposed model is trained on both the source dataset for the person re-identification task and the target dataset for the pedestrian detection task. To facilitate appearance and structure disentangling on target instances without identity information, we firstly apply the generative module to the source instances, followed by aligning the appearance code from different domains generated from the shared appearance encoder. This alignment facilitates to transfer appearance information from source instances to target instances. The pedestrian instances with appearance codes from the source dataset and structure codes from the target dataset can greatly augment the diversity of target instances. The generated pedestrian instances combined with reliable instances from the labeled and unlabeled images of the target domain are fed into the external classifier which is placed at the end of the generative module. The classifier returns integrity signals of given instances to the generative module to promote the discrimination of the generative module on partially-represented instances. The trained classifier is separated to generate pseudo annotations for unannotated images of the target dataset. Both annotated data and unannotated data with pseudo annotations are fed into the base detector to enhance its discrimination and localization. The whole framework of the generative module is shown in Fig. 1.

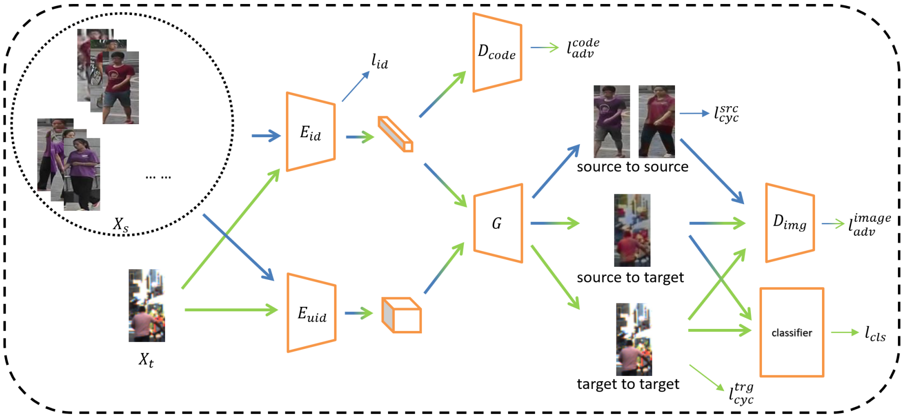

Fig. 1. An overview of the proposed framework. The generative module includes an id-related appearance encoder E_{id}, id-unrelated structure encoder E_{uid}, decoder G and image-level discriminator D_{img} shared across different domains. A code-level domain classifier D_{code} is introduced to compete with the id-related encoder for the alignment of id-related codes from different domains. An external classifier is introduced to train on appearance-rich instances and send integrity signals to the generative module to improve its discrimination on well/poorly represented instances. The different colored lines denote different dataflow to corresponding sub-modules and losses. Mixed colored lines denote the common dataflow of the source data and target data.

3.2 Information transferring from id-rich instances to id-absence instances

In our setting, the training data includes a small amount of annotated images $\mathbb{I}_a$ and a large amount of unannotated images $\mathbb{I}_u$ from the target scene. In addition, source instances with identities $(x_i^s, y_i^s)_{i=1}^N \in X_s$ are necessary to facilitate information disentangling, where N indicates the number of images, $y_i^s \in [1, K]$ indicates the identities of corresponding instances, and K indicates the total amount of identities in the source dataset. In the beginning, trustworthy target instances $x_i^t \in X_t$ are generated from $\mathbb{I}_a$ and $\mathbb{I}_u$ through an initial detector pretrained on $\mathbb{I}_a$. Our target is to transfer the appearance information from X_s to X_t while keeping the pedestrian shape in X_t. For this purpose, we should solve the problem of pedestrian disentangling and source-target domain discrepancy under the condition of lacking identity-level supervision and fully-represented instances in the target domain.

A shared id-related encoder $E_{id} : x_i \Rightarrow \upsilon_i$ and a shared id-unrelated encoder $E_{uid} : x_i \Rightarrow \omega_i$ are introduced to disentangle id-related appearance code and id-unrelated structure code from instances of different domains, and a shared decoder $G : (\upsilon_i, \omega_j) \Rightarrow x_{ij}$ is introduced to combine codes from different instances to generate unseen instances. To facilitate accurate disentangling, the generative module should be able to reconstruct an instance from the same structure code and the appearance codes of the same identities of itself at the source scene:

$$l_{recons} = \mathbb{E}[\left\|x_i^s - G(\upsilon_j^s, \omega_i^s)\right\|] \ . \tag{1}$$

where υ_j^s is an id-related appearance code from instance x_j^s with the same identity as x_i^s. In Eq. 1, $i = j$ is agreeable for self-reconstruction. For keeping id-related information into the id-related code, E_{id} should also learn to discriminate the identities of source instances:

$$l_{id} = \mathbb{E}[-\log(p(y_i^s|x_i^s))] \ . \tag{2}$$

where $p(y_i^s|x_i^s)$ is the predicted probability of x_i^s from E_{id}. Encouraging E_{id} to learn about the identity information facilitate the disentangling of id-related code. Further, code-level cycling supervision is essential for preventing information loss from disentangling:

$$l_{cyc}^{src} = \mathbb{E}[\left\|\upsilon_i^s - E_{id}(G(\upsilon_i^s, \omega_j^s))\right\| + \left\|\omega_j^s - E_{uid}(G(\upsilon_i^s, \omega_j^s))\right\|] \ . \tag{3}$$

In this case, υ_i^s and ω_i^s are well disentangled while together remaining all information of original instances.

Direct introduction of target instances can not help the information transferring from source instances to target instances because of the domain discrepancy. Although source instances and target instances share the id-related and id-unrelated encoder, the id-related encoder mainly accounts for the diversity of generative instances. In the experiments, we show that the domain adversarial training on the id-unrelated structure code overwhelms the information transferring on the id-related appearance code and further destroys the target style, like the background. A domain classifier D_{code} is introduced to match the distribution of source and target data at the id-related code-level based on the adversarial training with the id-related encoder. The domain-adversarial loss function l_{adv}^{code} is as follows:

$$l_{adv}^{code} = \mathbb{E}[\log D_{code}(E_{id}(x_i^s)) + \log(1 - D_{code}(E_{id}(x_j^t)))] \ . \tag{4}$$

To guarantee the information completion, code-level cycling supervision is also employed in the target instances, which is as follows:

$$l_{cyc}^{trg} = \mathbb{E}[\left\|\upsilon_i^t - E_{id}(G(\upsilon_i^t, \omega_j^t))\right\| + \left\|\omega_j^t - E_{uid}(G(\upsilon_i^t, \omega_j^t))\right\|] \ . \tag{5}$$

Further, to encourage the reality of produced images, an adversarial training is introduced to align the distribution of generated instances and real instances at the image-level of different domains, and the corresponding adversarial loss is as follows:

$$l_{adv}^{image} = \mathbb{E}[\log D_{img}(x_i) + \log(1 - D_{img}(G(\upsilon_i, \omega_j)))] \ . \tag{6}$$

where D_{img} denotes the image discriminator, and υ_i and ω_j are extract from x_i and x_j, which are drawn from the union of X_s and X_t. The image discriminator shared across different domains focuses on flaws of images generated from the id-related code and the id-unrelated code of different domains at the same time. The generalization of the image discriminator leads to the generalization of the generative module on different domains, which can reduce domain discrepancy implicitly. The adversarial training at the code-level and image-level jointly promotes encoders to learn domain-invariant features.

Finally, we append a classifier C at the end of the generative module to learn about the appearance-rich instances directly. We experimentally explore that the anchor-free CSP-based [22] detector, which makes a prediction based on whether each location is the center of each pedestrian, is inclined to generate a large number of easy positive and negative instances, which are harmful to achieving a robust classifier. We adopt the focal loss [19] on the binary classification to fit this problem, which is shown as follows:

$$l_{cls} = \mathbb{E}[-\alpha(1 - C(x_i))^{\gamma} \log(C(x_i))] \ . \tag{7}$$

where α and γ are focusing factors, and x_i is drawn from the union of X_t and the set of generated instances with appearance codes from source instances and structure codes from target instances. The introduction of appearance-rich instances from the generative module and high-confidence instances from unannotated images encourages the classifier to become more robust to the appearance variance and specific scene variance in the target domain. In return, the classifier sends integrity signals for given pedestrian instances to the generative module, encouraging the generative module to discriminate the well/poorly represented instances.

We jointly train the id-related appearance encoder, id-unrelated structure encoder, decoder, image-level discriminator, domain classifier and external classifier using a weighted summation of the losses as follows:

$$l_{all} = \lambda_{recons} l_{recons} + \lambda_{id} l_{id} \ + \lambda_{cyc}(l_{cyc}^{src} + l_{cyc}^{trg}) + l_{adv}^{code} + l_{adv}^{image} + l_{cls}. \tag{8}$$

where λ_{recons}, λ_{id} and λ_{cyc} are the weighting factor to control the contributions of the reconstruction item, identity discrimination item and code-level cycling supervision item, respectively. As the common setting in [12, 15, 43], we set $\lambda_{recons} = 2$ and $\lambda_{cyc} = 1$ to prevent the information from losing when disentangling, and set $\lambda_{id} = 0.5$ to avoid the negative effect of low-quality generated images at the beginning. For the classifier, we set $\gamma = 2$ and $\alpha = 0.25$ as suggested in [19].

3.3 Re-training detector

The introduction of appearance-rich information from X_s and diverse scene information from X_t promotes the classifier's ability to identify pedestrians from backgrounds. We adopt the trained classifier to generate pseudo annotations for the unannotated images, which are leveraged by the detector together with the limited annotated images. We adopt different strategies for learning different types of data.

In this work, We adopt a CSP-based detector D_{det} which consists of a ResNet-based [11] backbone and three heads for pedestrians' center recognition, height regression and height offset regression, respectively. Each mini-batch contains randomly sampled images from annotated images $\mathbb{I}_a$ and unannotated images $\mathbb{I}_u$ with pseudo annotations, which are generated from the initial detector followed by filtering through the trained classifier. The focal loss is used for binary center prediction, which is shown as follows:

$$l_{det} = -\sum_{u=1}^{W}\sum_{v=1}^{H} \beta_{uv}(1 - D_{det}(f(x_i)))^{\delta} \log(D_{det}(f(x_i))) \ . \qquad (9)$$

where x_i denotes images from either $\mathbb{I}_a$ or $\mathbb{I}_u$, $f(*)$ denotes concatenated features from several layers of the backbone, W and H denote the width and height of the concatenated feature maps. In addition, β_{uv} and δ denote the focusing factors, in which δ is constantly set to 2 and β_{uv} is gained as follows:

$$\beta_{uv} = \begin{cases} 1_{\{y_{uv}=1\}} + 1_{\{y_{uv}=0\}}(1 - M_{uv})^{\tau} & x_i \in \mathbb{I}_a \ , \\ 1_{\{y_{uv}=1\}}\nu & x_i \in \mathbb{I}_u \ . \end{cases} \qquad (10)$$

where y_{uv} is the label of each location in the feature map indicating whether the location is the center of each pedestrian, M_{uv} is a Gaussian-based mask centered at each object to weaken the ambiguity of background points around the center, τ is a penalty factor, and ν is a factor to control the contributions of the pseudo-annotated images. We set τ and ν to 4 and 1 in all experiments. For annotated images, we depress the contributions of locations around each object's center but keep other locations' contributions unchanged. For unannotated images, since there exists unavoidable missing on challenging pedestrians, the negatives are rejected while keeping only the contributions of positives for training.

Smoothing L1 loss is implemented for height regression and height offset regression, which is shown as follows:

$$l_{loc} = \sum \ \phi(p^{\xi}, g^{\xi}) \ . \qquad (11)$$

where

$$\phi(p^{\xi}, g^{\xi}) = \begin{cases} \frac{1}{2}\left|p^{\xi} - g^{\xi}\right|^{2} & if \left|p^{\xi} - g^{\xi}\right| < 1 \ , \\ \left|p^{\xi} - g^{\xi}\right| - \frac{1}{2} & otherwise \ . \end{cases} \tag{12}$$

In Eq. 12, $p = (p^h, p^o)$ represents the predicted height and the offset to the height and $g = (g^h, g^o)$ represents the closest ground-truth's height and offset to the height, where $g^o = (\frac{g^x}{r} - \lfloor\frac{g^x}{r}\rfloor, \frac{g^y}{r} - \lfloor\frac{g^y}{r}\rfloor)$, (g^x, g^y) is the center location of the corresponding ground truth, and r is the downsampling factor for the feature map.

The final optimization is formulated as follows:

$$\min_{\theta} \eta_c \mathbb{E}[l_{det}] + \eta_r \mathbb{E}[l_{loc}] \ . \tag{13}$$

where θ denotes the parameters of the base detector and η_c and η_r are weighting factors to control the relative contributions of the classification item and regression item.

4 Experiments

In this section, we verify the effectiveness of the proposed method in transferring appearance information from integral instances in a source domain to incomplete instances in a target domain. We adopt Market-1501 [40] as the source dataset and conduct the validation through experiments on three wide-used benchmark datasets: Caltech1X [7], CityPersons [39], and KITTI [9]. The appearance-enhanced pedestrian instances further improve the performance of the classifier, which is used to produce reliable pseudo annotations for unannotated images. The pseudo-annotated images and manually annotated images are dependable for re-training an improved detector. We manage to surpass the previous state-of-the-art methods over all benchmark datasets in the semi-supervised setting.

4.1 Experimental settings

In our semi-supervised setting, only a small portion of target images are annotated. Unless otherwise stated, 5% of training images are randomly sampled as fully annotated images and the remaining training images keep unannotated in the training stage. As annotations of KITTI's test set are not accessible, we randomly sample 2/3 of annotated images of the public training set as the training data and the remaining data are treated as validation data. For Caltech1X and CityPersons, we adopt annotated images of the public training set as the training data and verify the effectiveness of our proposed method on the public test set and validation set, respectively. Since different datasets adopt different protocols for evaluation, we follow the standard evaluation criterion of average-log Miss Rate (MR) [7], indicating the ratio of missing ground truths over all ground truths, for Caltech1X and CityPersons, and Average Precision (AP) [8], indicating the ratio of detected ground truths over the total detected boxes, for KITTI.

We follow the same network structure of DG-Net [41] with the proposed code-level domain classifier, composed of three fully-connected layers, for adversarial training and the VGGNet-based classifier, trained on the appearance-enhanced pedestrian instances, for delivering the signal of instances' completeness to the generative module. We implement the transferring task on datasets of different domains. The target datasets are composed of high-confidence but partially-represented pedestrian instances extracted from annotated images and unannotated images through the pre-training base detectors. The whole training stop at 100K iterations and each mini-batch includes 4 images from the source dataset and 4 images from the target dataset. Each image is scaled to 256 × 128. We adopt SGD to train E_{id} with a learning rate of 0.002 and momentum of 0.9, and Adam [14] to train E_{uid}, G, D_{code}, D_{img} with a learning rate of 0.0001 and set $\beta_1 = 0$, $\beta_2 = 0.999$. Another SGD optimizer is adopted to train the classifier with an initial learning rate of 0.001 and a momentum of 0.9. All learning rates are decreased by 10 times at 60K, 90K and 95K iterations. For the inference of the generative module, pedestrian instances from the source dataset and the target dataset with a uniform resolution of 256 × 128 are fed into our proposed framework under an inference speed of 87 ms per instance on the platform with one GTX 2080Ti.

We adopt CSP as our base detector with ResNet-50 [11] as the backbone and follow the official code with the original setting when pre-training the base detector on the annotated images. When re-training the base detector with annotated images and unannotated images with pseudo annotations, we set the $\eta_c = 0.01$ and $\eta_r = 0.1$ in Eq. 13 on three benchmark datasets. We use Adam to optimize the detector with initial learning rates of 10^{-4}, 2×10^{-4} and 2×10^{-4} and stop training at 120K, 120K and 160K iterations on Caltech1X, CityPersons and KITTI, respectively. We optimize the network on one GPU (GTX 2080Ti) with mini-batches of 8, 2 and 4 images on Caltech1X, CityPersons and KITTI, respectively. For the inference of the re-trained base detector, images in Caltech1X, CityPersons and KITTI, which are uniformly rescaled to resolutions of 480×640, 1024×2048, and 384×1280, respectively, are fed into the re-trained base detector under inference speeds of 41 ms per frame, 373 ms per frame, and 65 ms per frame on the platform with one GTX 2080Ti.

4.2 Evaluation of the Generative Ability

Our approach is built on the idea of disentangling and reconstructing instances from reliable to unreliable domains through adversarial training at image-level and code-level. To well reconstruct appearance information in the unreliable target domain in which instances are likely to be partial, we incorporate a classifier to backward integrity signals to the generative module, encouraging the generative module to recognize the body part in each target instance. We compare the proposed method with other representative methods, including DG-Net [41],

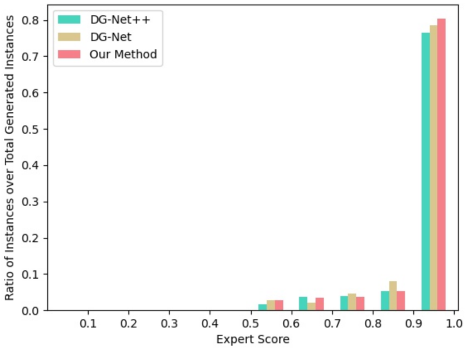

Fig. 2. The score distribution generated from the "expert" classifier over the generated pedestrian instances from DG-Net, DG-Net++, and our proposed method on KITTI.

DG-Net++ [44] and CycleGAN [43]. Since DG-Net can only work on single domains, we adapt DG-Net to be able to incorporate the instances from the target domain without identity information. As DG-Net++ is a multi-stage method for image transferring and person identification, we only adopt the results of the first stage which is the main stage for learning information transferring for comparison.

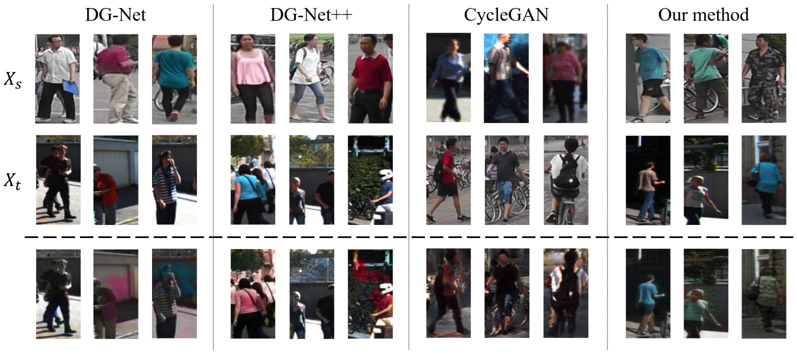

Fig. 3. Comparison of the generated instances from representative methods and our proposed method on KITTI.

We adopt a VGGNet-based expert model, which is trained on instances extracted from fully-annotated images, to evaluate the reality and reliability of the generated instances. Fig. 2 shows the confidence score distribution of the generated instances by DG-Net, DG-Net++ and our proposed method on KITTI. Compared to DG-Net and DG-Net++, the ratio of high confidence instances is larger. It means that the expert model highly appraises the generated instances and verifies the reliability of the generated instances.

We also compare the generated instances with the representative methods on KITTI. As shown in Fig. 3, CycleGAN can only transfer the whole style from the source domain to the target domain and cannot transfer the appearance-encoded information to the target instances. Without adversarial training at code-level, DG-Net fails to accurately locate the pedestrian body parts in the target instances, leading to unnaturally appearance inpainting. Although DG-Net++ can encode id-related codes from different domains into the same code space, the model cannot identify the body parts in the target instances and is inclined to overpaint the appearance information to the target instances. Our method can locate the body part in each target instance and inpaint the appearance information from the source instance into the correct location.

Fig. 4 shows examples of generated pedestrian instances of three benchmark datasets. Although some pedestrian instances from the target domains have incomplete body structures, the proposed model can still identify the body parts in the instances and inpaint the appearance information from the source instances into the body structures.

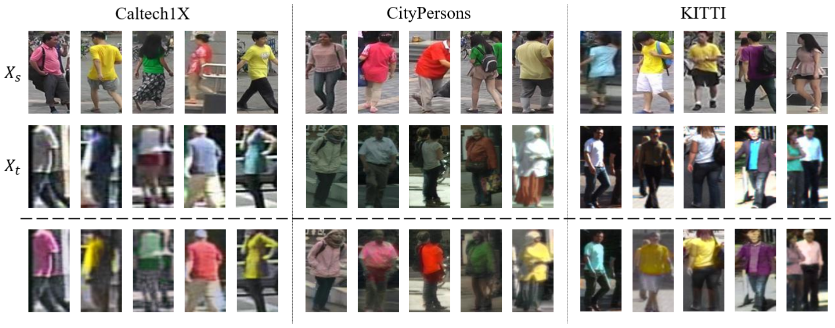

Fig. 4. Examples of pedestrian instances generated from our proposed method on Caltech1X, CityPersons, and KITTI.

Discussion. We claim that the id-related appearance code accounts mainly for the diversity when transferring between different domains. We also conduct an experiment to explore the case of deploying code-level domain classifiers to compete with both the id-related encoder and id-unrelated encoder. As shown in Fig. 5, the pedestrian instances generated from the generative module with separate code-level domain classifiers on both the id-related code and id-unrelated code are far from realistic. The competition between the id-unrelated encoder and the domain classifier is overwhelming, leading to the alignment of appearance codes of different domains and appearance information transferring more difficult. Therefore, an additional domain classifier for the id-unrelated encoder is helpless for generating appearance-enhanced pedestrian instances.

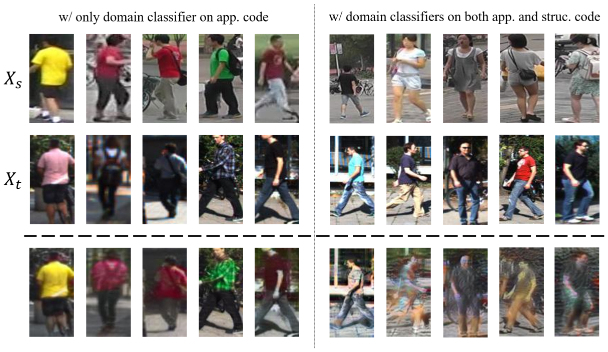
w/ only domain classifier on app. code
w/ domain classifiers on both app. and struc. code
X_s
X_t

Fig. 5. Examples of pedestrian instances generated from the generative module with different types of domain classifiers on KITTI.

4.3 Evaluation of the Discriminative Ability

We adopt CSP model as the base detector with ResNet-50 as the backbone network, and follow the experimental setting as [22] except for the incorporation of unannotated images with pseudo annotations. We firstly achieve a baseline result, termed as 'Base Detector (Ann-only)', from the base detector trained on annotated images only, which is treated as the lower bound for evaluation. We also build a model, called 'Base Detector (Ful-sup)', which is trained on fully-annotated images as the upper bound for evaluation. The comparison is shown in Table 1. We can observe that 'Base Detector (Ann-only)' is far from satisfactory due to the heavy reduction of annotated images. However, the joining of pseudo-annotated images promotes the performance by a large margin. In particular, the performance is promoted from 29.61 to 18.41, and 32.67 to 18.03 in MR for Caltech1X and CityPersons, respectively. Similarly, the incorporation of pseudo-annotated images advances the performance from 60.76 to 73.57 in AP for KITTI. Finally, we can observe that the performance in the semi-supervised setting is close to the performance of the model under sufficient supervision on KITTI.

We perform a comprehensive comparison with the existing state-of-the-art pedestrian detection methods in Table 1. Most methods are devised to achieve pedestrian detection with full annotations. For a fair comparison, these methods with open-source codes are trained on the annotated images as strong baselines. We can note that performances of methods that are initially designed to train on fully-annotated images, like Faster R-CNN [25], RPN+BF [38], ALFNet [21] and PDOE+RPN [42], deteriorate when given limited annotations and underperform our proposed method. Semi-supervised methods like SPV-RPN [35], PISD-RPN [33] and SaPA [36] outperform the supervised methods when given annotated and unannotated images. Our proposed method still surpasses SaPA, the current state-of-the-art method that worked in the semi-supervised setting, among three datasets. Especially, the proposed method achieves performance improvement by

Table 1. Evaluation results of the proposed method and competing methods on Caltech1X, CityPersons, and KITTI.

Methods	Caltech1X (MR↓)	CityPersons (MR↓)	KITTI (AP↑)
Faster R-CNN [25]	60.98	52.73	50.77
RPN+BF [38]	39.16	-	-
SDS-RPN [1]	35.66	49.52	-
ALFNet [21]	45.95	39.54	-
PDOE+RPN [42]	35.68	42.44	-
Variant SemiBoost [34]	52.53	-	-
SPV-RPN [35]	32.05	-	-
PISD-RPN [33]	23.79	-	-
SaPA [36]	18.55	26.54	67.97
Base Detector (Ann-only)	29.61	32.67	60.76
Base Detector (Re-trained)	18.41	18.03	73.57
Base Detector (Ful-sup)	11.40	12.02	78.78

about 8 percentage points and 6 percentage points on CityPersons and KITTI, respectively, and performs comparably with SaPA on Caltech1X.

5 Conclusion

In this work, we propose a framework that disentangles pedestrian instances into id-related appearance codes and id-unrelated structure codes from different domains which can provide pedestrian instances at different body-integrity levels. To reduce the discrepancy of appearance codes from different domains, a domain classifier is introduced to compete with the id-related appearance encoder. The generative module can capture integrity signals sent from an external classifier trained on the appearance-enhanced pedestrian instances, and inpaint the appearance factor into the remaining body structure of each instance in the target domain. The trained classifier has improved generalization to discriminate hard positive/negative and leads to better pseudo-annotations. Both the annotated images and pseudo-annotated images further promote the discrimination and localization of the re-trained detector.

Acknowledgements This work was supported in part by the Natural Science Foundation of Guangdong Province (Project No. 2020A1515010484, 2022A1515011160), in part by the National Natural Science Foundation of China (Project No. 62072189), and in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11201220).

References

1. Brazil, G., Yin, X., Liu, X.: Illuminating pedestrians via simultaneous detection & segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4950–4959 (2017)
2. Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: European conference on computer vision. pp. 354–370. Springer (2016)
3. Chen, Z., Ouyang, W., Liu, T., Tao, D.: A shape transformation-based dataset augmentation framework for pedestrian detection. International Journal of Computer Vision **129**(4), 1121–1138 (2021)
4. Cheung, E., Wong, A., Bera, A., Manocha, D.: Mixedpeds: Pedestrian detection in unannotated videos using synthetically generated human-agents for training. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018)
5. Chi, C., Zhang, S., Xing, J., Lei, Z., Li, S.Z., Zou, X.: Pedhunter: Occlusion robust pedestrian detector in crowded scenes. Proceedings of the AAAI Conference on Artificial Intelligence **34**(07), 10639–10646 (Apr 2020)
6. Chi, C., Zhang, S., Xing, J., Lei, Z., Li, S.Z., Zou, X.: Relational learning for joint head and human detection. Proceedings of the AAAI Conference on Artificial Intelligence **34**(07), 10647–10654 (Apr 2020)
7. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 304–311. IEEE (2009)
8. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). pp. 172–189 (2018)
13. Kim, J.U., Park, S., Ro, Y.M.: Robust small-scale pedestrian detection with cued recall via memory learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3050–3059 (2021)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European conference on computer vision (ECCV). pp. 35–51 (2018)
16. Li, C., Xu, T., Zhu, J., Zhang, B.: Triple generative adversarial nets. Advances in neural information processing systems **30** (2017)
17. Li, J., Liang, X., Shen, S., Xu, T., Feng, J., Yan, S.: Scale-aware fast r-cnn for pedestrian detection. IEEE transactions on Multimedia **20**(4), 985–996 (2017)

18. Lin, S., Wu, W., Wu, S., Xu, Y., Wong, H.S.: Unreliable-to-reliable instance translation for semi-supervised pedestrian detection. IEEE Transactions on Multimedia (2021)
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
20. Liu, S., Huang, D., Wang, Y.: Adaptive nms: Refining pedestrian detection in a crowd. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6459–6468 (2019)
21. Liu, W., Liao, S., Hu, W., Liang, X., Chen, X.: Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 618–634 (2018)
22. Liu, W., Liao, S., Ren, W., Hu, W., Yu, Y.: High-level semantic feature detection: A new perspective for pedestrian detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5187–5196 (2019)
23. Mao, J., Xiao, T., Jiang, Y., Cao, Z.: What can help pedestrian detection? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3127–3136 (2017)
24. Mhalla, A., Maamatou, H., Chateau, T., Gazzah, S., Amara, N.E.B.: Faster r-cnn scene specialization with a sequential monte-carlo framework. In: 2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 1–7. IEEE (2016)
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
26. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-supervised self-training of object detection models. In: 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1. vol. 1, pp. 29–36 (2005)
27. Wang, M., Wang, X.: Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In: CVPR 2011. pp. 3401–3408. IEEE (2011)
28. Wang, X., Wang, M., Li, W.: Scene-specific pedestrian detection for static video surveillance. IEEE transactions on pattern analysis and machine intelligence **36**(2), 361–374 (2013)
29. Wang, X., Hua, G., Han, T.X.: Detection by detections: Non-parametric detector adaptation for a video. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 350–357. IEEE (2012)
30. Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., Shen, C.: Repulsion loss: Detecting pedestrians in a crowd. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7774–7783 (2018)
31. Wu, J., Zhou, C., Zhang, Q., Yang, M., Yuan, J.: Self-mimic learning for small-scale pedestrian detection. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2012–2020 (2020)
32. Wu, J., Peng, Y., Zheng, C., Hao, Z., Zhang, J.: Pmc-gans: Generating multi-scale high-quality pedestrian with multimodal cascaded gans. arXiv preprint arXiv:1912.12799 (2019)
33. Wu, S., Lin, S., Wu, W., Azzam, M., Wong, H.S.: Semi-supervised pedestrian instance synthesis and detection with mutual reinforcement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5057–5066 (2019)
34. Wu, S., Wong, H.S., Wang, S.: Variant semiboost for improving human detection in application scenes. IEEE Transactions on Circuits and Systems for Video Technology **28**(7), 1595–1608 (2017)

35. Wu, S., Wu, W., Lei, S., Lin, S., Li, R., Yu, Z., Wong, H.S.: Semi-supervised human detection via region proposal networks aided by verification. IEEE Transactions on Image Processing **29**, 1562–1574 (2019)
36. Wu, W., Jiao, Q., Wong, H.S., Li, G., Wu, S.: Learning scene-adaptive pseudo annotations for pedestrian detection in semi-supervised scenarios. Knowledge-Based Systems **243**, 108439 (2022)
37. Zeng, X., Ouyang, W., Wang, M., Wang, X.: Deep learning of scene-specific classifier for pedestrian detection. In: European Conference on Computer Vision. pp. 472–487. Springer (2014)
38. Zhang, L., Lin, L., Liang, X., He, K.: Is faster r-cnn doing well for pedestrian detection? In: European conference on computer vision. pp. 443–457. Springer (2016)
39. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3221 (2017)
40. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person reidentification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. pp. 1116–1124 (2015)
41. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2138–2147 (2019)
42. Zhou, C., Yuan, J.: Bi-box regression for pedestrian detection and occlusion estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 135–151 (2018)
43. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
44. Zou, Y., Yang, X., Yu, Z., Kumar, B., Kautz, J.: Joint disentangling and adaptation for cross-domain person re-identification. In: European Conference on Computer Vision. pp. 87–104. Springer (2020)