

Semi-Supervised Pedestrian Instance Synthesis and Detection with Mutual Reinforcement

Si Wu^{1,2} Sihao Lin¹ Wenhao Wu¹ Mohamed Azzam² Hau-San Wong²
¹School of Computer Science and Engineering, South China University of Technology
²Department of Computer Science, City University of Hong Kong
 cswusi@scut.edu.cn, linsihao6@gmail.com, wenhaowu.chn@gmail.com
 m.azzam@my.cityu.edu.hk, cshswong@cityu.edu.hk

Abstract

We propose a GAN-based scene-specific instance synthesis and classification model for semi-supervised pedestrian detection. Instead of collecting unreliable detections from unlabeled data, we adopt a class-conditional GAN for synthesizing pedestrian instances to alleviate the problem of insufficient labeled data. With the help of a base detector, we integrate pedestrian instance synthesis and detection by including a post-refinement classifier (PRC) into a minimax game. A generator and the PRC can mutually reinforce each other by synthesizing high-fidelity pedestrian instances and providing more accurate categorical information. Both of them compete with a class-conditional discriminator and a class-specific discriminator, such that the four fundamental networks in our model can be jointly trained. In our experiments, we validate that the proposed model significantly improves the performance of the base detector and achieves state-of-the-art results on multiple benchmarks. As shown in Figure 1, the result indicates the possibility of using inexpensively synthesized instances for improving semi-supervised detection models.

1. Introduction

Pedestrian detection is a fundamental and critical step toward many real-world applications, such as surveillance and autonomous driving. Recent progress [1] [48] in pedestrian detection is mainly attributed to adapting deep convolutional neural networks (CNNs) to this task. However, there are still various challenges. Collecting and manually annotating a large amount of data for supervised learning requires lots of time and considerable human effort. In this paper, we limit our discussion to semi-supervised pedestrian detection. In our setting, there are a limited amount of labeled data and a large amount of unlabeled data. It is critical for our task to exploit the unlabeled data to facilitate learning on the

Figure 1. Pedestrian instance synthesis and detection on the CUHK-Square dataset. In a semi-supervised setting (5% labeled data), the proposed model generates high-fidelity pedestrian instances. When including the synthesized instances in the training of the PRC, the resulting PRC surpasses a base detector even with full supervision as shown in the right subfigure (lower is better).

labeled data. Although the state-of-the-art detection performance is promising, we found that it drops significantly as the amount of labeled data reduces. For instance, the model ‘RPN+BF’ [47] achieves a log-average miss rate of 9.6% in the FPPI (false positives per image) range $[10^{-2}, 10^0]$ on the Caltech1X benchmark [7], but the performance significantly degrades to 39.2% when only 5% of the total training data was used. In many real-world applications, insufficiency of labeled data often happens. Meanwhile, the performance of semi-supervised pedestrian detection is still far from being satisfactory.

Although some efforts have been made to combat the problem of insufficient labeled data, e.g., [38] [40] [41], most of them apply the current detector to collect new instances from unlabeled data and then re-train the detector. The main drawback of those methods is that the correctness of the collected instances cannot be guaranteed. Other works of solving the same problem is synthesizing pedestrian instances through rendering 3D human models [15] [4] [19]. However, the synthesized pedestrian in-

Figure 2. Illustration of the semi-supervised pedestrian instance synthesis and detection mechanism of our proposed GAN-based approach.

stances look unrealistic and unnatural. According to the study in [48], the performance of pedestrian detection methods heavily depends on the quality and diversity of training data. *How to readily exploit the available unlabeled data* is one of the most critical issues for semi-supervised pedestrian detection. Alternatively, it is promising to adopt generative adversarial networks (GANs) [13] to synthesize photo-realistic pedestrian instances. This makes us consider whether both pedestrian instance synthesis and detection could be taken into account simultaneously if we want to perform semi-supervised pedestrian detection.

The key idea of this work is mutual reinforcement between pedestrian instance synthesis and detection. Specifically, we adopt a Faster R-CNN [34] as our base detector. To develop a detection model with better generalization capability, we introduce an additional post-refinement classifier (PRC) following the Faster R-CNN. Inspired by TripleGAN [23], the PRC is incorporated with a generator to deal with the problem of insufficient labeled data. In addition, a class-conditional discriminator competes with both the generator and PRC in the minimax game. To fool this discriminator, the generator attempts to synthesize high-fidelity instances for each category, and the PRC tries to produce more accurate predictions on the unlabeled instances. To encourage the generator to synthesize more realistic pedestrian instances, an additional class-specific discriminator is included in our framework to focus on distinguishing the real and synthesized pedestrian instances. We also adopt a class-wise mean feature matching step to regularize the generator and alleviate the domain shift, such that the class-conditional distribution of the synthesis instances can match with that of the real pedestrian instances in the latent space learnt by PRC. Better classification can in turn lead to more accurate guidance to the generator. Consequently, the proposed model is able to improve both pedestrian instance synthesis and detection in the semi-supervised setting. To ensure test efficiency, the PRC is finally applied to per-

form pseudo-labeling on unlabeled data. We aim to re-train our base detector on the pseudo-labeled data, such that the resulting detector is expected to have similar performance to the PRC. Beyond our expectation, we found that the re-trained detector can achieve comparable or even superior results. An overview of the proposed approach is shown in Figure 2.

1.1. Contributions

The main contributions of the proposed approach can be summarized as follows:

- (1) We develop a new semi-supervised GAN-based framework, which effectively exploits unlabeled data for scene-specific instance synthesis with high-fidelity. This work provides new insights into semi-supervised pedestrian detection.
- (2) In our framework, the PRC, a generator and two kinds of discriminators can be trained jointly to facilitate mutual reinforcement between pedestrian synthesis and detection.
- (3) We further explore how to re-train the base detector using the pseudo-labeled data provided by the PRC. The resulting model can be used as a standalone detector which is capable of competing with or even outperforming the PRC without affecting efficiency.
- (4) We conduct thorough experiments to validate the effectiveness of the proposed approach. We show that our approach is effective in semi-supervised pedestrian instance synthesis. We also demonstrate that our approach significantly improves the performance of semi-supervised pedestrian detection.

To the best of our knowledge, this work is the first attempt to incorporate scene-specific pedestrian instance synthesis into the overall detection framework in a semi-supervised setting.

2. Related Work

With the recent development of CNNs, a remarkable progress has been made in the field of pedestrian detection. Since we adopt a Faster R-CNN [34] as our base pedestrian detector, we mainly review the recent works on CNN-based pedestrian detection methods. A number of widely used CNNs [21] [16] [36] have been applied to object detection and achieved a noticeable performance improvement as reported in previous works [34] [12] [6] [29]. Faster R-CNN has exhibited an impressive capability in general object detection. Based on that model, many pedestrian detection methods have been developed, and great progress has been achieved. For instance, Zhang et al. [47] replaced the downstream classifier of Faster R-CNN with a boosted forest model. When incorporating with a hard example mining strategy, their model improves detection performance significantly. Similarly, Hu et al. [18] incorporated the features extracted from the convolutional feature maps of a CNN into a boosted decision model. Instead of feeding the original images, You et al. [45] studied the mechanism of filtered channel features and extended it by using two or more convolutional layers with designed kernels upon HOG+LUV feature maps. The extended filtered channel features improved detection performance over the original ones.

To address the problems of large variance in pedestrian scale and occlusion, Cai et al. [3] proposed a multi-scale CNN to perform detection on multiple output layers. In [24], Li et al. proposed a scale-aware Fast R-CNN (SA-FastRCNN) model, in which multiple subnetworks are jointly trained to detect pedestrians with scales in different ranges. An active detection model (ADM) [50] could also provide more accurate prediction of multi-scale pedestrians by adopting a set of coordinate transformations with multi-layer feature representations. To simultaneously handle the two problems, Lin et al. [25] jointly trained a multi-scale network and a human parsing network. The former network learns multi-grained features which are useful for detecting small-size pedestrians, and the latter network learns a fine-grained attention map to improve the detection of occluded pedestrians according to the visible parts. On the other hand, Wang et al. [39] proposed a part and context network (PCN), which integrates complementary branches capturing semantic information of body part and contextual information to address the problem of occlusion. Another strategy is to apply a guided attention mechanism to focus on visible parts of occluded pedestrians. Zhang et al. [49] incorporated the Faster R-CNN with an add-on attention network (FasterRCNN+ATT). The attention mechanism across CNN channels was useful to reveal various occlusion patterns, such that the performance of occluded pedestrian detection was improved. In contrast, Wang et al. [42] focused on optimizing detection bounding box localization. In their model, a bounding box regression loss was included in the overall

loss function of Faster R-CNN to enhance pedestrian detection in crowd scenarios.

To improve classification on hard pedestrian/background instances, the fused deep neural network (F-DNN) [9] allows multiple parallel networks to refine the final prediction. To obtain context information, auxiliary segmentation tasks have been included in pedestrian detection models. For instance, Fidler et al. [10] applied the learnt segmentation masks to facilitate the detection task. In addition, segmentation results can be used to guide pedestrian detection as in [14]. In [2], a simultaneous detection and segmentation R-CNN (SDS-R-CNN) was proposed to improve pedestrian detection by including an auxiliary task of semantic segmentation. In addition, Costea et al. [5] performed semantic segmentation on channel feature maps to construct semantic channels, which can be viewed as additional visual cues. This consequently leads to performance gains. Further, Ouyang et al. [30] integrated feature extraction, deformation handling, occlusion handling and classification into a joint deep learning framework. In addition to enhancing pedestrian detection performance, there are also lots of effort made to speed up detection. YOLO [32] [33], SSD [26] and DSSD [11] were proposed to combine the region proposal generation and classification stages.

While significant effort has been devoted to improving the detection performance in the fully-supervised setting, the resulting improvement comes at the cost of the required huge amount of labeled data. Meanwhile, only a few of the existing works focus on studying semi-supervised pedestrian detection. To address this problem, a variant semi-supervised boosting model was proposed in [44]. They exploited the similarity between labeled data and unlabeled data in the process of training boosted models. Another relevant work is the adoption of a two-stage detection method in [43]. The authors applied a self-paced learning paradigm to progressively train an AlexNet to score proposals generated by an initial detector. Significantly different from these methods, our work is applied to semi-supervised pedestrian instance synthesis and detection. Instead of collecting possibly unreliable instances from unlabeled data for constructing additional labeled data, we improve model training by leveraging synthesized instances with high certainty. To mutually reinforce pedestrian instance synthesis and detection, we jointly train the corresponding fundamental networks in a minimax game.

3. Method

To facilitate semi-supervised pedestrian detection, our strategy is to exploit class-conditional GAN-based data augmentation to address the problem of insufficient scene-specific labeled data. We adopt a Faster R-CNN as our base detector and initially train it on the labeled data only. The input image goes through the backbone network,

and then proposals are generated by a region proposal network (RPN) on the resulting feature maps. To perform more accurate classification for the proposals, we build a semi-supervised GAN-based model, which consists of 4 fundamental networks: the PRC C , a class-conditional generator G , a class-conditional discriminator D_{con} , and a class-specific discriminator D_{spe} . These networks compete in a four-player minimax game.

3.1. Instance Synthesis and Classification in Semi-supervised Setting

Optimizing the generator. Let p_l and p_u denote the distributions of labeled and unlabeled data, respectively. $(x, y) \sim p_l$ represents a labeled data pair, where y is the class label of a sample x . Similarly, if $x \sim p_u$, this means that x represents an unlabeled sample. To make the synthesized pedestrian instances have a similar background environment as the real instances, we propose to train a class-conditional generator G , which translates pairs of random vector and class label (z, y) sampled from a prior distribution p_s to new scene-specific instances. In addition to the adversarial training term \mathcal{L}_{adv}^G , we include a mean feature matching term \mathcal{L}_{feaMat} into the overall loss function of G to alleviate the domain shift between real and synthesized data. Therefore, the optimization of the generator can be formulated as follows:

$$\min_G \mathcal{L}_{adv}^G + \mu \mathcal{L}_{feaMat}, \quad (1)$$

where

$$\begin{aligned} \mathcal{L}_{adv}^G = & E_{(z,y) \sim p_s} \log [1 - D_{con}(G(z, y), y)] \\ & + E_{(z,y) \sim p_s} \log [1 - D_{spe}(G(z, y), y^+)], \end{aligned} \quad (2)$$

y^+ denotes the label of the pedestrian class, and the weighting factor μ is used to adjust the relative importance between adversarial training and distribution matching. To fool D_{con} , G learns to synthesize instances of both pedestrian and background classes. To compete with D_{spe} , G needs to put more focus on pedestrian instance synthesis. As will be demonstrated by the experiments in Section 4.2 (Figure 4 and Table 1), combining D_{con} and D_{spe} indeed leads to higher synthesis quality of pedestrian instances. Furthermore, the mean feature matching term \mathcal{L}_{feaMat} is defined as follows:

$$\begin{aligned} \mathcal{L}_{feaMat} = & E_{(x,y) \sim p_l} \mathbb{1}^{y=y^+} f_C(x) \\ & - E_{(z,y) \sim p_s} \mathbb{1}^{y=y^+} f_C(G(z, y)), \end{aligned} \quad (3)$$

where the function $\mathbb{1}^{(\cdot)}$ return 1 if the input is true and 0 otherwise, and $f_C(\cdot)$ denotes the features on the last hidden

layer of the PRC. Minimizing this term encourages G to generate pedestrian instances that match the statistics of the real instances in the latent space, such that incorporation of the synthesized instances is effective for classifier training.

Optimizing the PRC. The goal of the proposed model is not just to synthesize more realistic pedestrian instances, but also improve the accuracy of pedestrian detection. Toward this end, both real and synthesized instances are used for training the PRC. In addition to the mean feature matching term in Eq.(3), the overall loss function of the PRC includes an adversarial training term \mathcal{L}_{adv}^C and a classification evaluation term \mathcal{L}_{claEva} . The corresponding optimization formulation is presented as follows:

$$\min_C \mathcal{L}_{adv}^C + \alpha \mathcal{L}_{feaMat} + \beta \mathcal{L}_{claEva}, \quad (4)$$

where α and β are weighting factors for controlling the relative importance of the corresponding terms in the overall loss function. For an unlabeled sample, let \hat{y} denote the estimate of its class label according to the prediction $p(x|C)$, where the PRC is parameterized by C . To compete with the discriminator D_{con} , the PRC needs to make \hat{y} as accurate as possible. Therefore, the adversarial training term \mathcal{L}_{adv}^C is defined as follows:

$$\mathcal{L}_{adv}^C = E_{x \sim p_u} p(x|C) \log [1 - D_{con}(x, \hat{y})]. \quad (5)$$

Since the synthesized instances are associated with specified class labels, they can be used for supervised learning (i.e. in the same way as the manually labeled instances being used). In addition, the PRC can also learn from the unlabeled real instances by including a conditional entropy term with respect to the posterior probability distribution into the overall loss function. Thus, the term \mathcal{L}_{claEva} is defined as follows:

$$\begin{aligned} \mathcal{L}_{claEva} = & E_{(x,y) \sim p_l} [-y \log p(x|C)] \\ & + E_{x \sim p_u} [-p(x|C) \log p(x|C)] \\ & + E_{(z,y) \sim p_s} [-y \log p(G(z, y)|C)], \end{aligned} \quad (6)$$

Minimizing the overall loss function in Eq.(6) forces the PRC to correctly classify both the labeled real data and the synthesized data, along with producing confident predictions on the unlabeled real data. With the inclusion of high-fidelity synthesized instances, the PRC can learn to generalize well to other unseen instances.

Adversarial training. We follow the adversarial training scheme of the Triple-GAN model in general. The discriminator D_{con} competes with both the generator and the PRC in a minimax game by distinguishing the labeled data pairs $\{(x, y) | (x, y) \sim p_l\}$ from two kinds of fake data pairs: $\{(G(z, y), y) | (z, y) \sim p_s\}$ and $\{(x, \hat{y}) | x \sim p_u\}$. We

can formulate the corresponding optimization problem as follows:

$$\begin{aligned} \max_{D_{\text{con}}} & E_{(x,y) \sim p_l} \log D_{\text{con}}(x, y) \\ & + \frac{1}{2} E_{(z,y) \sim p_s} \log (1 - D_{\text{con}}(G(z, y), y)) \\ & + \frac{1}{2} E_{x \sim p_u} \log (1 - D_{\text{con}}(x, \hat{y})) . \end{aligned} \quad (7)$$

Meanwhile, the other discriminator D_{spe} learns to further judge whether the pedestrian instance is real or fake. Its optimization formulation can be expressed as follows:

$$\begin{aligned} \max_{D_{\text{spe}}} & E_{(x,y) \sim p_l, y=y^+} \log D_{\text{spe}}(x, y) \\ & + E_{(z,y) \sim p_s, y=y^+} \log (1 - D_{\text{spe}}(G(z, y), y)) . \end{aligned} \quad (8)$$

D_{con} and D_{spe} play different roles in our model. The former makes G learn the background information as well as pedestrians, and the latter enforces G to synthesize more realistic pedestrian instances with better shapes and details. When we jointly train the four networks, the generator and the PRC are able to mutually reinforce each other in the form of generating more realistic pedestrian instances for data augmentation, along with producing more accurate guidance on category information. Therefore, the proposed model is able to facilitate both pedestrian instance synthesis and detection in the semi-supervised setting.

3.2. Enhancement of the Base Detector

The PRC can be reasonably expected to have a better capability of distinguishing unseen pedestrians from the background. However, the PRC cannot be applied alone to perform efficient pedestrian detection on full images. This is because there are a large number of bounding boxes in the background, which can be quickly filtered out by the pre-trained base detector. To improve the test efficiency, we aim to re-train the detector on the pseudo-labeled data, such that the new model is able to approximate the PRC in classification performance without affecting efficiency. Toward this end, we apply the pre-trained base detector and PRC to scan unlabeled images. The locations of the detection bounding boxes and the pedestrians in the corresponding regions are taken as the pseudo ground truths for the unlabeled images. These images, combined with automatically generated annotations, are used to re-train the base detector. Note that the pseudo-labeled images are partially labeled since some pedestrians may be missed by the detector. In the following, we describe in detail how to re-train the detection model on partially labeled data.

The training set contains a limited number of manually labeled images and a large number of pseudo-labeled images. In each iteration, the mini-batch holds two random-

ly selected images from these two kinds of data. To update the model, training samples are in the form of randomly selected candidate bounding boxes from a sliding window on a specific feature layer. Each sample x can be represented by the corresponding bounding box coordinate $b = (b^x, b^y, b^w, b^h)$, where (b^x, b^y) denotes the top-left corner, b^w denotes the width, and b^h denotes the height. The corresponding label y indicates whether b is a pedestrian bounding box. A sample is considered as positive if the intersection-over-union (IoU) ratio of the corresponding bounding box and the closest ground-truth bounding box is greater than 0.5, and negative otherwise. The detection loss function det for each training sample is made up of the category prediction component and position regression component, defined as follows:

$$\text{det}(x, y) = -y \log p(x | \mathcal{R}) + r 1^{\{y=y^+\}} \text{locReg}(b, \tilde{b}), \quad (9)$$

where \mathcal{R} denotes the parameters of the detection model, \tilde{b} represents the ground-truth bounding box closest to the proposal b , and r is a weighting factor of the regression term locReg , which is defined as follows:

$$\text{locReg}(b, \tilde{b}) = \frac{\|b - \tilde{b}\|}{\{x, y, w, h\}}, \quad (10)$$

where $\|\cdot\|$ denotes the robust L_1 loss function used in Fast R-CNN. Note that this term works only for the positive training samples as $y = y^+$ in that case. Re-training the base detector on both real-labeled and pseudo-labeled data can be expressed through the following optimization formulation:

$$\begin{aligned} \min_{\mathcal{R}} & E_{(x,y) \sim p_l} \text{det}(x, y) + E_{(x,y) \sim p_l, y=y^+} \text{det}(x, y) \\ & + c E_{(x,y) \sim p_l, y=y^+ \text{ or } (x,y) \sim p_l, y=y^-} \frac{\|p(x | \mathcal{R}) - p(x | \mathcal{R})\|}{2}, \end{aligned} \quad (11)$$

where y^- denotes the pseudo labels of an instance x collected from the pseudo-labeled data, and p_l represents the corresponding distribution. The third term in Eq.(11) is included to encourage the detection model to produce consistent predictions for a sample x located in the neighborhood $N(x)$ centered at x (e.g., $\text{IoU}(x, x') > 0.7$). Both c and r are weighting factors for adjusting the relative contribution of the pseudo-labeled samples and the consistency regularization term.

4. Experiments

In this section, we focus on verifying that our proposed model for pedestrian instance synthesis significantly improves the accuracy of pedestrian detection in the semi-supervised setting. In order to conduct such validation, we

(a) Real pedestrian instances

(b) Synthesized pedestrian instances

Figure 3. Examples of the real pedestrian instances and synthesized pedestrian instances produced by the proposed model on CUHK-Square (top row), MIT-Traffic (middle row) and Caltech1X (bottom row).

use three benchmark datasets: MIT-Traffic [38], CUHK-Square [37] and Caltech-USA [7]. In the experiments, we managed to achieve a significant improvement over the baseline detector and outperform the previous state-of-the-art methods on all the test datasets. In particular, the performance of our approach is comparable/superior to those of the fully supervised models on both MIT-Traffic and CUHK-Square.

4.1. Experimental Setting

In all the experiments, our semi-supervised setting is that only 5% of the training images are fully annotated, and the remaining 95% images are regarded as unlabeled data without including any annotations in the training process. We follow the evaluation criterion in [8], in which the heights of pedestrians are at least 50 pixels and visual levels are at least 65%. The detection methods are evaluated using the standard log-average miss rate (MR) which is the official metric of Caltech-USA dataset. The average value is computed at 9 FPPI rates in the log-space range $[10^{-1}, 10^0]$ ($[10^{-2}, 10^0]$ for Caltech1X), which is the same as the main competing methods. For our base detector Faster R-CNN, we directly use the source code provided by the authors. We implement the proposed semi-supervised GAN-based model using TensorFlow. The model is trained by using the Adam solver [20]. For re-training the base detector, we use the stochastic gradient descent optimizer with a momentum of 0.9. We start with an initial learning rate of 10^{-3} , and then decrease it by 10 times after 2000 (20000 for Caltech1X) iteration.

4.2. Evaluation of the Synthesis Quality

First of all, we investigate the effectiveness of the proposed semi-supervised GAN-based model in pedestrian instance synthesis. Figure 3 shows examples of synthesized instances on the three test datasets. The synthesized pedes-

Table 1. Comparison of the proposed model and its variants in terms of quality of synthesis on Caltech1X.

Method	IS	FID
Ground-truth	3.05 ± 0.35	-
SN-GAN [28]	1.89 ± 0.05	216.66
Our Model w/o D_{spe}	2.39 ± 0.13	103.60
Our Model	2.74 ± 0.08	44.18

(a) SN-GAN (b) Ours w/o D_{spe} (c) Ours

Figure 4. Comparison of the proposed model and its variants in pedestrian instance synthesis on Caltech1X.

trian instances have a complete body structure and look natural with satisfactory qualities. Compared to the real pedestrian instances, the synthesized ones also have a similar style and degree of clarity. This indicates that the generator in our model can effectively capture the scene information and generate reasonable scene-specific instances.

Quality of synthesis. To highlight the advantage of combining a class-conditional discriminator and a class-specific discriminator, we compare the proposed model with a state-of-the-art GAN model, SN-GAN [28], trained on labeled data only. We also compare with a variant ‘Our Model w/o D_{spe} ’ by removing the class-specific discriminator from our model. Figure 4 shows additional synthesis results for comparing these three models on Caltech1X. We can make the following observations: the synthesized

Figure 5. The score distributions of the ‘expert’ classifier for the synthesized pedestrian instances produced by the proposed model and variants on Caltech1X.

instances of both ‘Our Model w/o D_{spe} ’ and ‘Our model’ have substantially higher quality than those of SN-GAN. In addition, ‘Our model’ performs better than ‘Our Model w/o D_{spe} ’ in reducing the appearance ambiguity and preserving more reasonable structure of the human body. Further, we evaluate these models in terms of inception score (IS) [35] and Fréchet inception distance (FID) [17] in Table 1. We can clearly observe that ‘Our model’ achieves the highest value of IS and lowest value of FID on Caltech1X.

Expert evaluation. Our final objective is to improve semi-supervised pedestrian detection via the inclusion of synthesized pedestrian instances. IS and FID cannot guarantee the semantics of synthesized instances. To handle this problem, we propose to adopt a fully supervised classification network as an ‘expert’ to score the synthesized pedestrian instances. We consider that the confidence score of the ‘expert’ can indicate whether the pedestrians in the synthesized images are well represented. Figure 5 shows the score distribution of the synthesized instances produced by SN-GAN, ‘Our model w/o D_{spe} ’ and ‘Our model’ on Caltech1X. Compared with SN-GAN and ‘Our model w/o D_{spe} ’, there are much more synthesized pedestrian instances of ‘Our model’ that have high confidence scores (>0.5). On the other hand, we apply PCA to visualize the distributions of the synthesized pedestrian instances by SN-GAN and ‘Our model’. Each instance is represented by the features extracted from the last hidden layer of the ‘expert’ network. As shown in Figure 6, we can notice that the synthesized pedestrian instances of ‘Our model’ match well with the real pedestrian instances than those of SN-GAN. This is important for pedestrian instance augmentation. When we include our synthesized data in the training process, both the amount and diversity of pedestrian instances can significantly increase while reducing the risk of misleading the PRC.

4.3. State-of-the-Art Comparison

In this subsection, we perform a comparison with the state-of-the-art semi-supervised pedestrian detection methods on the test datasets. Our base detector is a Faster R-CNN with VGG-16 [36] as the backbone network, and the

(a) SN-GAN (b) Our Model

Figure 6. The embedding of the real pedestrian instances and synthesized pedestrian instances on Caltech1X. PCA is adopted to project the features extracted from the last hidden layer of the ‘expert’ network to 2D.

Table 2. The log-average miss rates of our model and the competing methods in the FPPI range $[10^{-1}, 10^0]$ on CUHK-Square and MIT-Traffic.

Method	CUHK-Square	MIT-Traffic
Generic Detector Adaptation [38]	0.8240	0.7915
Transferring Boosted Detector [31]	0.6936	0.6770
Confidence-encoded SVM [41]	0.6352	0.6475
Transferring Attributes [51]	0.6249	-
Data-reconstructed CNN [46]	0.5361	0.5327
SMC Faster R-CNN [27]	0.4326	0.4703
Variant SemiBoost [44]	0.4290	0.3647
Temporal Ensembling [22]	0.2820	0.4494
Self-paced CNN [43]	0.2742	0.2687
Our Base Detector (Initial)	0.3467	0.3458
Our Base Detector (Re-trained)	0.1924	0.1509

setting of the corresponding hyper-parameters is the same as [47]. We initially train the base detector on the labeled data only as the baseline, and the corresponding test results are used as the lower bound for our evaluation. We also report the results of the base detector retrained on the augmented labeled data through pseudo-labeling unlabeled data. Table 2 shows the detection results of our proposed approach and the competing methods on CUHK-Square and MIT-Traffic. The performance of ‘Our Base Detector (Retrained)’ is far better than ‘Our Base Detector (Initial)’. The performance gains reach about 15 percentage points on CUHK-Square and 19 percentage points on MIT-Traffic. The proposed approach outperforms the second-best method ‘Self-paced CNN’ by about 8 and 12 percentage points on these two datasets, respectively.

In addition, we also test our approach on a more complex dataset, Caltech1X. Previous works focus on performing fully supervised learning on this dataset. Consequently, we compare with the representative pedestrian detection models, e.g., the original Faster R-CNN, RPN+BF and SDS-RPN. These models are trained on the labeled data only. Different from other competing methods, ‘Variant SemiBoost’ is a non-deep semi-supervised method, which is trained on the same labeled data and unlabeled data as the

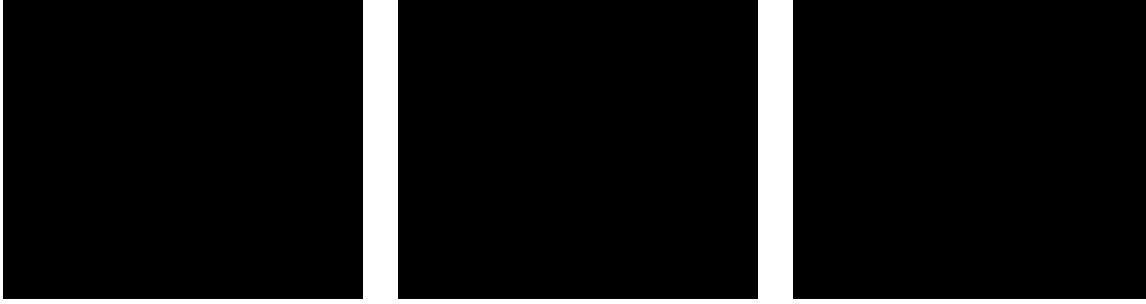


Figure 7. Comparison of the proposed approach with its variants on semi-supervised pedestrian detection on CUHK-Square, MIT-Traffic and Caltech1X. The log-average miss rate is computed in the FPPI range $[10^{-1}, 10^0]$ ($[10^{-2}, 10^0]$ for Caltech1X).

Table 3. The log-average miss rates of our model and the competing methods in the FPPI range $[10^{-2}, 10^0]$ on Caltech1X.

Method	Caltech1X
Faster R-CNN [34]	0.6098
RPN+BF [47]	0.3916
SDS-RPN [2]	0.3566
SDS-R-CNN [2]	0.3403
Variant SemiBoost [44]	0.5253
Our Base Detector (Initial)	0.4565
Our Base Detector (Re-trained)	0.2379

proposed approach. Table 3 shows the detection results of our approach and the competing methods. ‘Variant SemiBoost’ performs poorly. The proposed approach improves the baseline by about 22 percentage points and achieves the best result. This improvement is notable given this limited number of labeled images.

4.4. Discussion

To obtain a better insight into the effect of semi-supervised pedestrian instance synthesis and detection, we conduct more experiments in this subsection. Specifically, we demonstrate the performance of the PRC in our model on the test datasets to illustrate the benefits of synthesized pedestrian instances. We also train our base detector based on full supervision as ‘Our Base Detector (Ful-Sup)’, where all the training images are fully annotated. The detection-error-tradeoff curves of our model and its variants are plotted in Figure 7. The PRC significantly outperforms ‘Our Base Detector (Initial)’ in all the cases, which indicates that the synthesized instances are indeed useful for pedestrian instance augmentation. It is worth noting that the PRC even surpasses ‘Our Base Detector (Ful-Sup)’ on CUHK-Square. Compared to the PRC, ‘Our Base Detector (Re-trained)’ is able to achieve comparable performance on CUHK-Square and MIT-Traffic, and better performance on Caltech1X, which verifies that our re-training strategy is effective. Further, we present two ablation studies on Caltech1X to highlight the importance of mean feature match-

ing in our model. We build two variant models: ‘PRC w/o Syn. Ins.’ and ‘PRC w/o Fea. Mat.’. The former does not use the synthesized instances for training the PRC, and the latter disables the mean feature matching term $feaMat$ in Eq.(4). We can find that the two variants have a similar performance. Although they outperform the baseline, the improvement is not as significant as our full model. Therefore, we conclude that mean feature matching is an effective way to mitigate the domain shift and plays an important role in our semi-supervised GAN-based model.

5. Conclusion

In this paper, we explored how to synthesize scene-specific instances using GANs to address the problem of insufficient labeled data in semi-supervised pedestrian detection. Different from previous works on collecting new instances from unlabeled data, our approach addresses this issue via simultaneous pedestrian instance synthesis and improvement in classification. Toward this end, with the help of a base detector, we developed a semi-supervised GAN-based model to mutually reinforce a generator and the PRC. We also verified that the proposed model is capable of generating high-fidelity pedestrian instances with limited supervision. It was demonstrated that those instances indeed lead to significant performance gains in pedestrian detection on multiple datasets. Encouraged by the results, we anticipate that the proposed approach can be applied to other general object detection problems.

Acknowledgments

This work was supported in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11300715), in part by the National Natural Science Foundation of China (Project No. U1611461), in part by City University of Hong Kong (Project No. 7005055), in part by the Natural Science Foundation of Guangdong Province (Project No. 2016A030310422), and in part by the Fundamental Research Funds for the Central Universities (Project No. 2018ZD33).

References

- [1] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In *Proc. European Conference on Computer Vision*, pages 613 – 627, 2014.
- [2] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating pedestrians via simultaneous detection and segmentation. In *Proc. IEEE International Conference on Computer Vision*, pages 4960 – 4969, 2017.
- [3] Zhaowei Cai, Quanfu Fan, Rogerio S. Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *Proc. European Conference on Computer Vision*, pages 354 – 370, 2016.
- [4] Ernest Cheung, Anson Wong, Aniket Bera, and Dinesh Manocha. MixedPeds: pedestrian detection in unannotated videos using synthetically generated human-agents for training. In *Proc. AAAI Conference on Artificial Intelligence*, 2018.
- [5] Arthur Daniel Costea and Sergiu Nedevschi. Semantic channels for fast pedestrian detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2360 – 2368, 2016.
- [6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: object detection via region-based fully convolutional networks. In *Proc. Advances in Neural Information Processing Systems*, 2016.
- [7] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: a benchmark. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 304 – 311, 2009.
- [8] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: an evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743 – 761, 2012.
- [9] Xianzhi Du, Mostafa EL-Khamy, Jungwon Lee, and Larry S. Davis. Fused DNN: a deep neural network fusion approach to fast and robust pedestrian detection. In *Proc. IEEE Winter Conference on Applications of Computer Vision*, pages 953 – 961, 2017.
- [10] Sanja Fidler, Roozbeh Mottaghi, Alan Yuille, and Raquel Urtasun. Bottom-up segmentation for top-down detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3294 – 3301, 2013.
- [11] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Amrith Tyagi, and Alexander C. Berg. DSSD: deconvolutional single shot detector. In *arXiv preprint arXiv:1701.06659*, 2016.
- [12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 580 – 587, 2014.
- [13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Advances in Neural Information Processing Systems*, 2014.
- [14] Bharath Hariharan, Pablo Arbelaez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *Proc. European Conference on Computer Vision*, pages 297 – 312, 2014.
- [15] Hironori Hattori, Vishnu Naresh Boddeti, Kris Kitani, and Takeo Kanade. Learning scene-specific pedestrian detectors without real data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 770 – 778, 2016.
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. Neural Information Processing Systems*, 2017.
- [18] Qichang Hu, Peng Wang, Chunhua Shen, Anto van den Hengel, and Fatih Porikli. Pushing the limits of deep CNNs for pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(6):1358 – 1368, 2018.
- [19] Shiyu Huang and Deva Ramanan. Expecting the unexpected: training detectors for unusual pedestrians with adversarial in-posters. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [20] Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representations*, 2015.
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Neural Information Processing Systems*, pages 1106 – 1114, 2014.
- [22] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *Proc. International Conference on Learning Representations*, 2017.
- [23] Chongxuan Li, Kun Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. In *Proc. Advances in Neural Information Processing Systems*, pages 1195 – 1204, 2017.
- [24] Jianan Li, Xiaodan Liang, Shengmei Shen, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Scale-aware fast R-CNN for pedestrian detection. *IEEE Transactions on Multimedia*, 20(4):985 – 996, 2018.
- [25] Chunze Lin, Jiwen Lu, and Jie Zhou. Multi-grained deep feature learning for robust pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology (Early Access)*, 2018.
- [26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot MultiBox detector. In *Proc. European Conference on Computer Vision*, 2016.
- [27] Ala Mhalla, Houda Maamatou, Thierry Chateau, Sami Gazzah, and Najoua ESSOUKRI BEN Amara. Faster R-CNN scene specialization with a sequential Monte-Carlo framework. In *Proc. International Conference on Digital Image Computing: Techniques and Applications*, 2016.
- [28] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *Proc. International Conference on Learning Representation*, 2018.

- [29] Junhyug Noh, Soochan Lee, Beomsu Kim, and Gunhee Kim. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [30] Wanli Ouyang, Hui Zhou, Hongsheng Li, Quanquan Li, Junjie Yan, and Xiaogang Wang. Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1874 – 1887, 2018.
- [31] Junbiao Pang, Qingming Huang, Shuicheng Yan, Shuqiang Jiang, and Lei Qin. Transferring boosted detectors towards viewpoint and scene adaptiveness. *IEEE Transactions on Image Processing*, 20(5):1388 – 1400, 2011.
- [32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: unified, real-time object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [33] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [34] Shanqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. Advances in Neural Information Processing Systems*, 2015.
- [35] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Proc. Neural Information Processing Systems*, pages 2234 – 2242, 2016.
- [36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale recognition. In *Proc. International Conference on Learning Representations*, 2015.
- [37] Meng Wang, Wei Li, and Xiaogang Wang. Transferring a generic pedestrian detector towards specific scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3274 – 3281, 2012.
- [38] Meng Wang and Xiaogang Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3401 – 3408, 2011.
- [39] Shiguang Wang, Jian Cheng, Haijun Liu, Feng Wang, and Hui Zhou. Pedestrian detection via body part semantic and contextual information with DNN. *IEEE Transactions on Multimedia*, 20(11):3148 – 3159, 2018.
- [40] Xiaoyu Wang, Gang Hua, and Tony X. Han. Detection by detection: non-parametric detector adaptation for a video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 350 – 357, 2012.
- [41] Xiaogang Wang, Meng Wang, and Wei Li. Scene-specific pedestrian detection for static video surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):361 – 374, 2014.
- [42] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: detecting pedestrians in a crowd. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [43] Si Wu, Shufeng Wang, Robert Laganiere, Cheng Liu, Hau-San Wong, and Yong Xu. Exploiting target data to learn deep convolutional networks for scene-adapted human detection. *IEEE Transactions on Image Processing*, 27(3):1418 – 1432, 2018.
- [44] Si Wu, Hau-San Wong, and Shufeng Wang. Variant Semi-Boost for improving human detection in application scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(7):1595 – 1608, 2018.
- [45] Mingyu You, Yubin Zhang, Chunhua Shen, and Xinyu Zhang. An extended filtered channel framework for pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 19(5):1640 – 1651, 2018.
- [46] Xingyu Zeng, Wanli Ouyang, Meng Wang, and Xiaogang Wang. Deep learning of scene-specific classifier for pedestrian detection. In *Proc. European Conference on Computer Vision*, pages 472 – 487, 2014.
- [47] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is Faster R-CNN doing well for pedestrian detection? In *Proc. European Conference on Computer Vision*, pages 443 – 457, 2016.
- [48] Shanshan Zhang, Rodrigo Benenson, Mahamed Omran, Jan Hosang, and Bernt Schiele. Towards reaching human performance in pedestrian detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):973 – 986, 2018.
- [49] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in CNNs. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [50] Xiaowei Zhang, Li Cheng, Bo Li, and Hai-Miao Hu. Too far to see? not really! - pedestrian detection with scale-aware localization policy. *IEEE Transactions on Image Processing*, 27(8):3703 – 3715, 2018.
- [51] Xu Zhang, Fei He, Lu Tian, and Shengjin Wang. Cognitive pedestrian detector: adapting detector to specific scene by transferring attributes. *Neurocomputing*, 149:800 – 810, 2015.