Box Re-Ranking: Unsupervised False Positive Suppression for Domain Adaptive Pedestrian Detection

Weijie Chen$^{1,2}$, Yilu Guo$^2$, Shicai Yang$^2$, Zhaoyang Li$^2$, Zhenxin Ma$^2$, Binbin Chen$^2$, Long Zhao$^2$, Di Xie$^2$, Shiliang Pu$^2$, Yueting Zhuang$^1$

$^1$Zhejiang University, Hangzhou China
$^2$Hikvision Research Institute, Hangzhou, China

{chenweijie5,guoyilu5,yangshicai,lizhaoyang5,mazhenxin}@hikvision.com
{chenbinbin8,zhaolong11,xiedi,pushiliang.hri}@hikvision.com,yzhuang@zju.edu.cn

Abstract

False positive is one of the most serious problems brought by agnostic domain shift in domain adaptive pedestrian detection. However, it is impossible to label each box in countless target domains. Therefore, it yields our attention to suppress false positive in each target domain in an unsupervised way. In this paper, we model an object detection task into a ranking task among positive and negative boxes innovatively, and thus transform a false positive suppression problem into a box re-ranking problem elegantly, which makes it feasible to solve without manual annotation. An attached problem during box re-ranking appears that no labeled validation data is available for cherry-picking. Considering we aim to keep the detection of true positive unchanged, we propose box number alignment, a self-supervised evaluation metric, to prevent the optimized model from capacity degeneration. Extensive experiments conducted on cross-domain pedestrian detection datasets have demonstrated the effectiveness of our proposed framework. Furthermore, the extension to two general unsupervised domain adaptive object detection benchmarks also supports our superiority to other state-of-the-arts.

1. Introduction

Pedestrian detection [28, 6, 5, 32, 27, 31] is a very important problem in the field of computer vision and sees many practical applications, such as video surveillance, autonomous driving and so on. Generally, a well-performed pedestrian detection model is trained on a very large-scale annotated images (source domain data), and then embedded in edge devices to process the images sampled from practical scenarios (target domain data). This paradigm makes it a very challenging problem due to agnostic domain shift, namely different context between source domain and target domain. False positive is one of the most serious problems during cross-domain pedestrian detection. Two examples are shown in Fig.1. The left one is from social news, where a dog is false-detected as a pedestrian. The right one is an indoor video surveillance scenario for monitoring the break-in of strangers. However, what alerting is always a crowd of goldfishes raised in the room, making the users very annoying. Actually, false positive samples are totally diverse in different target domains. However, it is impossible to label each target domain data manually, which motivates us to study unsupervised false positive suppression for cross-domain pedestrian detection.

As mentioned above, a very large-scale source annotated data is a prerequisite to pre-train a well-performed pedestrian detection model. However, during domain adaptive optimization, such large-scale source data is hard to store in edge devices due to data transmission as well as data privacy-protection. Especially, source data is unable to suppress the unseen false positive, which is an open-set problem and can only find the answer from target domain. In the right image of Fig.1 detected by YOLOv3 [16] trained on MS COCO dataset [21], goldfish is never seen in MS COCO dataset, which is the main reason leading to false positive. Therefore, source domain data is inaccessible in this paper. Only a pre-trained pedestrian detection model and unlabeled target data collected online are provided for...
unsupervised false positive suppression.

The most challenging problem lies in no labels are available in target domain. The only supervision cue is the detected bounding boxes attached with their classification confidence from a source domain pre-trained model. Sorting the confidence from higher to lower scores, an object detection task can be viewed as a ranking task, which aims to optimize the positive and negative boxes in a good order. In other words, the occurrence of false positive actually behaves as a wrong order with true positive (or false negative). In this way, a false positive suppression problem can be modeled into a box re-ranking problem elegantly.

Empirically, the boxes with extremely high positive confidence are likely to be real positive boxes, especially when pre-trained model is trained on a very large-scale dataset collected from numerous source domains. These boxes can provide prior knowledge about true positive and act as seed positive boxes to rectify the order of the remaining ones. Initially, the seed positive boxes and the remaining ones are directly labeled as positive and negative for fine-tuning, and then the confidence of the remaining ones is re-predicted for re-ranking. Note that treating the remaining boxes as negative is a very critical step in our framework to suppress false positive. The positive candidates in the remaining boxes tend to achieve higher confidence than others after training. Recursively, the top-K positive candidates are removed from the remaining boxes and join into the positive counterpart to repeat the training process. We name the entire repetition as box re-ranking. To reduce the complexity, deep transfer clustering [17] is the first time to use in an object detection task and pack the boxes of the entire dataset with similar semantic information into a cluster. The boxes in a cluster are re-ranked as a whole, simplifying the granularity of box re-ranking from an instance level to a cluster level. Also, since pure seed positive boxes are the preliminaries to rectify the order of the remaining ones, label denoising for seed positive boxes is also an extra benefit from deep transfer clustering.

Another question followed by box re-ranking is when to terminate the repetition without using labeled validation data. Since we hope to keep the detection capacity of true positive unchanged after false positive suppression, we propose box number alignment, a self-supervised evaluation metric, for cherry-picking during box re-ranking. Specifically, the detected box number should keep unchanged before and after box re-ranking. Actually, in many practical scenarios, the false positive are usually occurred along with false negative, which means the more false positive are suppressed, the more false negative will be mined back. Overall, the entire pipeline of our proposed framework is shown as Fig.2.

Extensive experiments are conducted on four pedestrian detection adaptation tasks built with five public pedestrian detection datasets. To further verify our proposed method, we collect a more challenging pedestrian detection dataset for false positive suppression. Also, to compare with other related work, we extend our method to two general domain adaptive object detection tasks. Comprehensive experimen-
tal results and the corresponding analysis demonstrate the effectiveness of our proposed framework. It can suppress false positive by a large margin in an unsupervised way without declining the detection capacity of true positive. In summary, our main contributions are listed as follows:

- To the best of our knowledge, this is the first work aiming to suppress false positive for domain adaptive pedestrian detection in an unsupervised way.
- An object detection task is modeled into a ranking task innovatively, and a false positive suppression problem is turned into a box re-ranking problem elegantly, making it feasible to solve in an unsupervised way. Box number alignment, a self-supervised evaluation metric, is proposed for cherry-picking during box re-ranking.
- Five datasets are constructed to validate the effectiveness of our proposed unsupervised false positive suppression framework, one of which is our self-collected dataset. The extension to two general object detection adaptation tasks also support our superiority to other state-of-the-arts.

2. Unsupervised False Positive Suppression

It is extremely difficult to assign an exactly right pseudo label to each box in an unsupervised settings. To bypass this challenge, we model a false positive suppression problem into a box re-ranking problem. Also, to prevent the degeneration of detection capacity, box number alignment is proposed as a self-supervised evaluation metric to cooperate with box re-ranking.

2.1. Prerequisites

Before introducing our proposed unsupervised false positive suppression framework, three important prerequisites have to emphasize:

- The bounding boxes generated by the pre-trained model are directly applied as ground-truth for box regression training, no matter positive or negative boxes, since the location error is much weaker than classification error in object detection tasks as proved by [1].
- Reliable positive boxes are the seed prerequisite for our proposed framework to drive box re-ranking. Usually, those with extreme high positive confidence predicted from source domain pre-trained model are very likely to be real positive boxes. This condition is usually satisfied especially when the pre-trained model is trained on a very large-scale dataset collected from numerous source domains. Beyond this, deep transfer clustering is adopted for further label denoising and keeping the purity of seed positive boxes.
- In most practical scenarios, false positive are occurred along with false negative. After our proposed box re-ranking, the confidence position among false positive and false negative are swapped, which indirectly suppresses false positive while mining back a corresponding part of false negative.

2.2. Box Re-Ranking

As mentioned above, we select seed positive boxes with very high confidence (greater than a given threshold $h$, 0.95 by default in this paper), while the remaining ones are directly treated as negative, which can be formulated as:

$$\{ (b, c) \}_{N} \rightarrow \mathcal{X}$$

where $M_{S}$ is a given pre-trained model and $\mathcal{X}$ denotes target training dataset. Exploiting $M_{S}$ on $\mathcal{X}$, we can initially achieve bounding box set $\{ (b, c) \}_{P}$ where $b$ denotes a detected box attached with corresponding confidence $c$. $\mathcal{Y}$ is the initialized label set including positive part $P$ and negative part $N$ divided from $\{ (b, c) \}$ by a very high confidence threshold $h$.

$$M_{T} \leftarrow L_{\text{conf}}(M_{S}|\mathcal{X}; \mathcal{Y}) + \lambda L_{\text{loc}}(M_{S}|\mathcal{X}; \mathcal{Y})$$

(3)

Using the initial pseudo labels $(\mathcal{X}, \mathcal{Y})$, the source domain pre-trained model $M_{S}$ is fine-tuned into $M_{T}$ as shown in Eqn.3. The objective loss function is a weighted sum of the confidence loss (conf) and the localization loss (loc). As mentioned in 2.1, we directly use initially-generated boxes as ground-truth for box regression training. Usually, if the positive part is pure enough, the false positive will be well-suppressed. But it is quite easy to encounter the degeneration of true positive. To avoid this situation, more positive boxes should be mined back from the negative part elaborately.

$$\{ (b, c) \}_{T} = M_{T}(\mathcal{X})$$

(4)

Adopting the optimized model $M_{T}$, the pre-generated bounding boxes $\{ b, c \}$ are re-predicted to update their confidence. To prevent the catastrophic forgetting of bounding box localization, the ground-truth of localization always keeps unchanged. In this way, we use the initial bounding box $\{ (b, c) \}$ to match the re-predicted ones $\{ (b, c) \}_{T}$. Let $d_{ij} = \{ 1, 0 \}$ be an indicator for matching the $i$-th initial bounding box to the $j$-th re-predicted box. $d_{ij} = 1$ only if the IoU of this pair is the greatest one and the IoU should be greater than a given threshold (0.3 by default in this paper). So the confidence update process is:

$$c_{i} = \sum_{j} d_{ij} \cdot c_{t,j} \text{where } c_{t,j} \in \{ (b, c) \}_{t} \in \{ (b, c) \}_{T}$$

(5)
After confidence update, the bounding box \( \{b\} \) are re-ranked from higher to lower scores. The boxes in negative part with similar appearance to positive part will tend to achieve higher positive confidence \( N_{top} \). This process is actually similar to PU-Learning [3]. We update the label \( Y \) by removing \( N_{top} \) from \( N \) and adding \( N_{top} \) to \( P \):

\[
P = P \cup N_{top} \\
N = N \setminus N_{top} \\
Y = \{P, N\}
\]  

(6)

After updating \( Y \), Eqn.3-5 are repeated recursively until achieving the best performance.

### 2.3. Box Packing via Deep Transfer Clustering

It is too redundant to re-rank the boxes in an instance level, since no matter true positive and false positive usually appear frequently in a single target scene. To further discretize the granularity of box re-ranking, the boxes with similar semantic feature are packed into a cluster via deep transfer clustering:

\[
\min_{C \in \mathbb{R}^{d \times k}} \frac{1}{N} \sum_{n=1}^{N} \min_{y_n \in \{0,1\}^k} ||f_s(b_n) - C y_n||_2^2 \text{ s.t. } y_n^T 1_k = 1
\]  

(7)

where \( f_s(b_n) \) means we map the center of the \( n \)-th box to the feature map \( f_s \) generated by \( M_s(\cdot) \) and extract the corresponding feature of the box for clustering. \( f_s(b_n) \in \mathbb{R}^d \) and \( d \) is the channel size of \( f_s \). \( C \in \mathbb{R}^{d \times k} \) is the cluster centroid matrix, \( k \) is the cluster number and \( y_n \) is the cluster assignment for box \( b_n \). \( N \) is the box number of the entire dataset.

As shown in Fig.3, we use YOLOv3 [16] as an example since we mainly use YOLOv3 in this paper. We concatenate the last second layer of three FPN heads [20] for feature extraction, and directly map the center of the boxes to the corresponding positions in the feature map. By default, we use \( k \)-means for boxes clustering. Without specific statement, we over-cluster the boxes into 100 clusters in this paper. Note that the extracted features include semantic as well as scale and aspect ratio information, since they are used for classification and box regression. Scale and aspect ratio information is beneficial for clustering, since the boxes with similar semantic information usually share similar scale and aspect ratio.

After clustering, the boxes in the same cluster are re-ranked together as a whole via averaging the confidence in a cluster.

\[
c_i = \frac{1}{|\Omega|} \sum_{n \in \Omega} c_n \text{ where } \Omega = \{j\}, y_j = y_i
\]  

(8)

Concretely speaking, the boxes whose averaging confidence in the corresponding cluster is greater than \( h \) are selected as seed positive boxes. During box re-ranking, the candidate positive boxes are added from the negative part to the positive part cluster by cluster.

An extra benefit of deep transfer clustering is label denoising. A reliable assumption is that the boxes with similar feature should share a similar confidence score consistently. Clustering makes the confidence of each box more robust via voting, avoiding accidently high confidence brought by domain noise and achieving more reliable seed positive boxes. The outlier boxes of each cluster are annotated as ignored labels during optimization.

### 2.4. Self-Supervised Evaluation

In an unsupervised setting, no ground-truth labels are available in target data, which means no annotated validation data is available for cherry-picking during box re-ranking. Actually, this is also the main challenge of all unsupervised learning tasks. In this paper, we will introduce a self-supervised evaluation metric box number alignment to be an alternative of annotated validation data for false positive suppression tasks.

A main objective during false positive suppression is that the detection capacity of true positive should not be declined. Hence, we aims at keeping the number of detected boxes over a confidence threshold unchanged before and after box re-ranking. Actually, in practical application, a confidence threshold \( o \) is usually pre-set to output the detected boxes (Note that \( o \) is usually smaller than \( h \). In practical, \( o \) is usually set according to the False Positive Per Image on source domain validation set.). So the box number in the unlabelled validation set are countable and can be used as an evaluation metric for unsupervised false positive suppression. It can be formulated as:

\[
\min_i \text{ s.t. } C(M_{s,j}(\mathcal{X}), o) \geq C(M_s(\mathcal{X}), o)
\]  

(9)

where \( M_s \) and \( M_{s,j} \) denote the pre-trained model in source domain and the \( i \)-th optimized model in target domain during box re-ranking. \( \mathcal{X} \) is an unlabelled validation set in target domain, and \( o \) is a confidence threshold to count the output box number after NMS [24]. \( C(\cdot) \) separately represent the function of box counting. In general, with our proposed box number alignment, the more false positive are suppressed, the more false negative are mined back.
Unsupervised Domain Adaptive Object Detection It has made a great progress in the challenging unsupervised domain adaptation problem recently [26, 36, 34, 29, 30, 9]. In the object detection task, most of related methods delve into how to achieve cross-domain alignment between source domain and target domain with different solutions, such as DA-Faster [4], SW-Faster [19], Region-level Alignment [37], CR-DA-DET [33], ATF [12], style transfer based method [13] and so on. Different from these work, we model the object detection task into a box ranking task to solve the problem of unsupervised false positive suppression without using any domain alignment technique.

Positive-Unlabeled Learning As mentioned above, domain adaptive object detection is more likely an open-set problem, since the negative instances could be anything around the world, while the positive ones are unique with only domain shift. Therefore, it is naturally connected to positive-unlabeled learning [3]. Our proposed box re-ranking framework can be regarded as an improved version of positive-unlabeled learning to solve unsupervised false positive suppression problem. To the best of our knowledge, this is the first time we connect positive-unlabeled learning to domain adaptive object detection problem.

### 4. Results

#### 4.1. Experimental Settings

Datasets and network architecture. To demonstrate the effectiveness of our proposed method, we build four pedestrian detection domain adaptation tasks. MS COCO [21] is a very large-scale object detection dataset consisting of pedestrian detection, which is taken as source domain
dataset in this paper. Following YOLOv3 [16], we train a YOLOv3 608×608 with backbone Darknet-53 on MS COCO as a strong pre-trained model for unsupervised optimization on downstream target domains. Four challenging pedestrian detection datasets are selected as downstream target domains, including Caltech [8], Cityperson [35], KITTI [10, 2] and KAIST [14]. Briefly speaking, Caltech contains 10 hours of video from an urban driving environment, Cityperson records street views across 18 different cities in Germany with various weather conditions, KITTI is a popular urban object detection dataset, and KAIST is a multispectral pedestrian detection dataset with well-aligned visible / thermal pairs captured during day and night. Detailed information about training size, validation size, and image resolution on source and target domains are shown in Table 1. To further verify our proposed method, we also build a more challenging target domain dataset via self-collected data, and we plan to release this dataset later.

**Evaluation metrics.** To describe the performance of false positive suppression, log-average Miss Rate over False Positive Per Image (FPPI) ranging in $[10^{-2}, 10^0]$, which is short for MR in this paper, is utilized as a main evaluation metric. It is also popularly used in many other pedestrian detection work. A good order among true positive boxes and false positive boxes leads to a small MR. Without specific statement, we only determine MR under IoU=0.5. To further evaluate our proposed method comprehensively, Average Precision (AP) is also presented along with MR as an auxiliary evaluation metric.

**4.2. Implementation details**

We initialize $h = 0.95$ and $o = 0.4$ to drive seed positive boxes selection and *box number alignment* in a very conservative way. We carry out deep transfer clustering to the boxes with confidence greater than 0.1. Detaiely, the boxes in the clusters with average confidence greater than 0.95 are set as seed positive boxes. Note that we only carry out deep transfer clustering once before *box re-ranking*. In each repetition for confidence re-prediction, we zero out the confidence of initial boxes generated by the pretrained model, and match them to the re-predicted boxes generated by the optimized model with maximum IoU (at least greater than 0.3) for confidence updating. Without specific statement, in each repetition, the boxes in top-3 clusters with highest confidence in the negative part are mined back to positive part to drive the next repetition until reaching the condition of *box number alignment*. The experiments on following cross-domain pedestrian detection datasets share the same implementation configurations. The entire pipeline can be referred to Algorithm 1.

**4.3. Ablation Studies**

**Deep transfer clustering on cross-domain object detection task.** Despite agnostic domain shift, the boxes with similar semantic feature are able to get into the same cluster. What’s more, as what we claim above, the scale and aspect ratio information encoded into YOLOv3 feature for box regression are also beneficial for object clustering. As shown in Fig.4, the boxes with similar semantic, scale and aspect ratio information are well-grouped, which divides true positive and false positive boxes into different clusters.

**Performance in each repetition during box re-ranking.** To take a deep look at how *box re-ranking* works, we present the varying curve of MR as *box re-ranking* goes on. As shown in Fig.5, MR decreases at first since the or-
der of false positives with higher confidence are swapped with false negatives with lower confidence. As box re-ranking goes on, fewer and fewer false negatives are left to be swapped with false positives, resulting in that false positives might be added to positive part for fine-tuning and further leading to MR increasing. This is usually the extremely worst case resulted from over box re-ranking and violates the third prerequisite proposed in section 2.1. Our proposed box number alignment (short for BNA) is a useful self-supervised evaluation metric to avoid this situation. We terminate box re-ranking when the box number predicted by the optimized model is comparable as before under a given confidence threshold, namely \(\theta\) in this paper. \(\theta\) is usually used in practical object detection scenarios to output boxes. This operation can prevent our optimized model from over box re-ranking as well as from detection degeneration of true positive. To test BNA to the most extent, we venture to set \(\theta = 0.4\) to a very low score in this paper.

**Self-supervised vs. supervised evaluation.** We discuss what are the best performances our proposed box re-ranking can achieve via supervised evaluation with labeled validation data (best). And we also compare the performances achieved via BNA with supervised evaluation counterparts. As shown in Table 2, our method can decrease MR and increase AP by a large margin compared with source only, the performance directly tested by the pre-trained model. Also, the performance gap between supervised and self-supervised evaluation are relatively small.

**Comparison with supervised training.** We use ground-truth labels of training dataset on target domain for supervised fine-tuning. It can reflect the performance upper-bound (oracle) for the unsupervised optimization methods, so as to more precisely evaluate our proposed method. As shown in Table 2, our method can improve the performance in terms of MR and AP. Also, as shown in Table 6, our method can also settle the adaptation problem from RGB image to thermal image without effort. In general, there is still exist a performance gap between our method and oracle one. And we mainly impute it to the mis-detection of false negatives instead of the detection of false positives. And our paper only focus on false positive suppression. Nevertheless, based on our proposed false positive suppression framework, we hope to dig out more false negatives which are hard to detect even if we set the confidence threshold near to zero. And we leave this as future work.

**Instance-based vs. cluster-based box re-ranking.** We also carry out an ablation experiment on Caltech dataset to demonstrate the effectiveness of deep transfer clustering in our framework. As shown in Table 7, cluster-based box re-ranking performs much better than instance one. The main reason is that the noisy labels in seed positive boxes are cleaned via clustering and the confidence determined in a cluster is more robust for box re-ranking.

**A more challenging dataset.** The existing public pedestrian datasets are not specifically designed for cross-domain false positive suppression task. To further verify our method, we collect a more challenging dataset in an indoor environment. In this dataset, the false positives are actually goldfish raised indoor as shown in Fig. 8, since fish is an unseen object in MS COCO dataset. This is the reason why we call cross-domain false positive suppression is an open-set problem in the above section. As shown in Fig. 7, the detected goldfish averagely has 0.77 confidence to false-detect into a pedestrian (the highest confidence even goes to 0.99). Since few pedestrians appear in this dataset, we mix it with Caltech dataset (Caltech-Fish dataset) to conduct the following experiment. As shown in Tab. 8 and Fig. 7, our method can well-solve this problem. The confidence of detected goldfish is averagely lower to 0.01, while detection capacity of true positive is unchanged.

### 4.4. Discussion

**Visualization of false positive suppression.** Our method claims to re-rank the confidence order of false positive and false negative. In this way, with comparable box number to be detected, we can suppress the false positive while mining back the considerable part of false negative. As shown in Fig. 9, we can effectively suppress false positive especially those just produced by agnostic domain noise with different semantic information from pedestrians. Also, it is worth mentioning that our method can also rectify the bounding boxes with greater IoU matched with ground truth boxes, which may be owe to the Law of Large Numbers.

**Training cost.** We try to keep the training configuration exactly the same so as to compare the performance among different repetitions fairly during box re-ranking. We always initialize the model with pre-trained weights from source domain, and fine-tune the model on target domain by only 5 epochs, since no source domain data join training and thus avoiding domain conflict between source and tar-

| Method          | Reasonable MR | Reasonable AP | All MR | All AP |
|-----------------|---------------|---------------|--------|--------|
| Source only     | 20.06         | 79.83         | 65.20  | 41.10  |
| Oracle          | 11.25         | 88.18         | 52.47  | 55.22  |
| Ours (BNA)      | 14.76         | 85.37         | 60.54  | 46.41  |
| Ours (Best)     | 13.57         | 86.34         | 59.89  | 46.72  |

Table 2. Results of adaptation from COCO to Caltech.
| Method      | Reasonable | Bare | Partial | Heavy |
|-------------|------------|------|---------|-------|
|             | MR         | AP   | MR      | AP    | MR   | AP   | MR   | AP   |
| Source only | 22.50      | 91.95| 12.19   | 94.68 | 24.80| 88.26| 84.77| 24.60|
| Oracle      | 13.01      | 95.66| 8.28    | 96.52 | 12.75| 94.77| 45.27| 77.34|
| Ours (BNA)  | 17.04      | 93.81| 8.60    | 96.09 | 19.59| 91.10| 78.84| 31.97|
| Ours (Best) | 14.83      | 93.95| 7.80    | 96.25 | 16.87| 91.88| 77.77| 33.13|

Table 3. Results of adaptation from COCO to Cityperson.

**Figure 6.** Performance comparison before and after box re-ranking. (miss rate vs. FPPI.)

| Method      | Easy | Moderate | Hard |
|-------------|------|----------|------|
|             | MR   | AP       | MR   | AP   | MR   | AP   |
| Source only | 38.31| 51.09    | 50.37| 49.11| 60.14| 42.69|
| Oracle      | 19.77| 79.10    | 28.55| 73.67| 36.03| 69.41|
| Ours (BNA)  | 34.26| 57.14    | 46.46| 55.50| 56.85| 49.55|
| Ours (Best) | 30.04| 64.15    | 41.85| 61.13| 52.97| 53.88|

Table 4. Results of adaptation from COCO to KITTI.

| Method      | Day | Night |
|-------------|-----|-------|
|             | MR  | AP    | MR  | AP  |
| Source only | 74.33| 34.00  | 58.45| 50.35|
| Oracle      | 28.03| 82.05  | 11.63| 92.81|
| Ours (BNA)  | 57.11| 53.69  | 37.64| 70.66|
| Ours (Best) | 54.29| 57.88  | 32.43| 75.52|

Table 5. Results of adaptation from COCO to KAIST RGB.

| Method      | Reasonable | All |
|-------------|------------|-----|
|             | MR         | AP  | MR  | AP  |
| Ours-cluster (Best) | 13.57| 86.34 | 59.89| 46.72|
| Ours-instance (Best) | 15.12| 85.19 | 61.60| 44.60|

Table 7. Performance comparison between instance-based and cluster-based box re-ranking on Caltech.

Future work. Box number alignment in this paper is proposed for false positive suppression. Beyond false positive suppression, more effective unsupervised evaluation metrics should be developed in the future for unsupervised learning with less prerequisite dependency.
Figure 8. More strongly hard false positive examples are collected for false positive suppression in cross-domain pedestrian detection. Red, blue and green box separately denote ground-truth, false positive, and true positive box.

| Method            | AP of Car |
|-------------------|-----------|
| Source only       | 36.4      |
| Oracle            | 58.5      |
| DA-Faster [4]     | 38.5      |
| SW-Faster [19]    | 37.9      |
| Noise Labeling [18]| 43.0     |
| DA-Detection [13] | 43.9      |
| AT-Faster [12]    | 42.1      |
| Ours              | 44.2      |

Table 9. Results of adaptation from KITTI to Cityscapes.

| Method            | AP of Car |
|-------------------|-----------|
| Source only       | 33.7      |
| Oracle            | 58.5      |
| DA-Faster [4]     | 38.5      |
| Noise Labeling [18]| 43.0     |
| AT-Faster [12]    | 42.1      |
| Ours              | 44.0      |

Table 10. Results of adaptation from Sim10k to Cityscapes.

4.5. Extension to General Benchmarks

Since cross-domain unsupervised false positive suppression is a new task we propose in this paper, no previous work is appropriate to discuss with our proposed box re-ranking framework quantitatively. To further verify our proposed method, we also compare our proposed method with other state-of-the-arts in two general domain adaptive object detection datasets, including the adaptation from KITTI to Cityscapes [7] and the adaptation from Sim10k [15] to Cityscapes. Following other work [4, 19, 18, 13, 12], these two adaptation datasets use Faster-RCNN [25] for car detection and take Average Precision (AP) as the main evaluation metric. Different from YOLOv3, we use ROI pooling feature of each box for deep transfer clustering in Faster-RCNN. As shown in Table 9 and 10, our proposed method outperforms other competing methods.

5. Conclusion

In this paper, a new task, named unsupervised false positive suppression for domain adaptive pedestrian detection, is proposed. In order to bypass this challenge elegantly, we model an object detection task into a box ranking task among true positive and false positive, and further transform false positive suppression problem into a box re-ranking problem. Under this modeling, false positive is able to be suppressed effectively without any manual annotation, which has been supported by extensive experiments carried out in this paper. Our proposed framework is very practical to upgrade object detection model into scene-specific one. We hope our method can bring rich inspirations to the community of unsupervised learning.

References

[1] Ali Borji and Seyed Mehdi Iranmanesh. Empirical upper bound in object detection and more. arXiv:1911.12451, 2019.
[2] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. 2016.
[3] Elkan Charles and Noto Keith. Learning classifiers from only positive and unlabeled data. In KDD, 2008.
[4] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. V. Gool. Domain adaptive faster r-cnn for object detection in the wild. In CVPR, 2018.
[5] Zhou Chunluan and Yuan Junsong. Bi-box regression for pedestrian detection and occlusion estimation. In ECCV, 2018.
[6] Lin Chunze, Lu Jiwen, Wang Gang, and Zhou Jie. Graininess-aware deep feature learning for pedestrian detection. In ECCV, 2018.
[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In CVPR, 2016.
[8] Piotr Dolla´r. Pedestrian detection: A benchmark. 2009.
[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. JMLR.org, 2017.
[10] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. International Journal of Robotics Research, 32(11):1231–1237, 2013.
Figure 9. Visualization of false positive suppression (testing under IoU = 0.5). Top left: MS COCO to Caltech. Top right: MS COCO to cityperson. Bottom: MS COCO to KAIST Termal. Red, green and blue boxes denote ground-truth, true positive and false positive, respectively.

[11] Dayan Guan, Xing Luo, Yanpeng Cao, Jiangxin Yang, Yan-long Cao, G Vosselman, and Michael Ying Yang. Unsupervised domain adaptation for multispectral pedestrian detection. 2019.

[12] Z. He and L. Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In ECCV, 2020.

[13] H.-K Hsu, C.-H. Yao, Y.-H. Tsai, W.-C. Hung, H.Y. Tseng, M. Singh, and M.-H. Yang. Progressive domain adaptation for object detection. In WACV, 2020.

[14] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In CVPR, 2013.

[15] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan. Driving in the matrix: Can virtual worlds replace humangenerated annotations for real world tasks? In ICRA, 2017.

[16] Redmon Joseph and Farhadi Ali. Yolov3: An incremental improvement. arXiv:1804.02767, 2018.

[17] Han Kai, Vedaldi Andrea, and Zisserman Andrew. Learning to discover novel visual categories via deep transfer clustering. In ICCV, 2019.

[18] M. Khodabandeh, A. Vahdat, M. Ranjbar, and W. G. Macready. A robust learning approach to domain adaptive object detection. In ICCV, 2019.

[19] Saito Kuniaki, Ushiku Yoshitaka, Harada Tatsuya, and Saenko Kate. Strong-weak distribution alignment for adaptive object detection. In CVPR, 2019.

[20] Tsung Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In CVPR, 2017.

[21] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. 2014.

[22] Lihang Liu, Weiyao Lin, Lisheng Wu, Yong Yu, and Michael Ying Yang. Unsupervised deep domain adaptation for pedestrian detection. In ECCV, 2016.

[23] Kieu My, Andrew D. Bagdanov, Bertini Marco, and Alberto del Bimbo. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In ECCV, 2018.

[24] A. Neubeck and L. Gool. Efficient non-maximum suppression. 2006.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell., 39(6):1137–1149, 2017.

[26] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In CVPR, 2018.

[27] Zhang Shanshan, Yang Jian, and Schiele Bernt. Occluded pedestrian detection through guided attention in cnns. In CVPR, 2020.

[28] Song Tao, Sun Leiyu, Xie Di, Sun Haiming, and Pu Shiliang. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In ECCV, 2018.

[29] Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino. Adversarial feature augmentation for unsupervised domain adaptation. In CVPR, 2018.

[30] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. Neurocomputing, 312(OCT.27):135–153, 2018.

[31] Liu Wei, Liao Shengcai, Ren Weiqiang, Hu Weidong, and Yu Yinan. High-level semantic feature detection: A new perspective for pedestrian detection. In CVPR, 2020.

[32] Huang Xin, Ge Zheng, Jie Zequn, and Yoshie Osamu. Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In CVPR, 2020.

[33] C. Xu, X. Zhao, X. Jin, and X. Wei. Exploring categorical regularization for domain adaptive object detection. In CVPR, 2020.

[34] Ganin Yaroslav and Lempitsky Victor. Unsupervised domain adaptation by backpropagation. In ICML, 2015.

[35] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In CVPR, 2017.

[36] Lu Zhihe, Yang Yongxin, Zhu Xiatian, Liu Cong, Song Yizhe, and Xiang Tao. Stochastic classifiers for unsupervised domain adaptation. In CVPR, 2020.

[37] X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin. Adapting object detectors via selective cross-domain alignment. In CVPR, 2019.
Figure 10. More visualization of false positive suppression on public-released datasets (IoU=0.5), including KAIST, KITTI, Cityperson and Caltech. Under the constraint of our proposed box number alignment, the more false positives are suppressed, the more false negatives will be mined back. Red, blue and green box separately denote ground-truth, false positive, and true positive box. A relatively small part of blue boxes are resulted from mis-annotation of true positives.