

Unsupervised Deep Domain Adaptation on People Detection

Anonymous ECCV submission

Paper ID ***

Abstract. This paper addresses the problem of unsupervised domain adaptation on the task of people detection in crowded scenes. That is, given a deep detection model well-trained on source domain, we adapt it into scene-specific detectors for any target domain on which no annotations are available. Firstly, we utilize iterative algorithm to iteratively auto-annotate target samples with high confidence on people instance as training set for scene-specific model on target domain. However, auto-annotated samples not only are lack of negative samples, but contains false positive samples. Therefore, on the one hand, we reuse negative samples from source domain to compensate for imbalance between the amount of positive samples and negative samples. On the other hand, we design an unsupervised regularizer based on deep network to mitigate influence from data error. Besides, we transform the last full connected layer into two sub-layers— element-wise multiply layer and sum layer, on which the unsupervised regularizer can be added on. In experiments on people detection, the proposed method boosts recall by nearly 30% while precision stays almost the same. Furthermore, we perform our method on standard domain adaptation benchmarks on both supervised and unsupervised settings and our results are state of the art.

Keywords: Unsupervised Domain Adaptation, Unsupervised Regularizer, Deep Neural Network, People Detection

1 Introduction

Deep neural network has shown great power on traditional computer vision tasks, however, the labelled dataset should be large enough to train a deep model. In famous challenges such as PASCAL VOC and MS COCO, millions of labelled images are needed for training. This is also the case in surveillance applications. The annotation process for the task of people detection in crowded scenes is even more resource consuming, cause we need to label concrete locations of people instances. In modern society, there are over millions of cameras deployed for surveillance. However, these surveillance situations vary in lights, background, viewpoints, camera resolutions and so on. Directly utilizing models trained on old scenes will results in poor performance on the new situations due to data distribution changes. It is also unpractice to annotate people instances for every surveillance situation.

When there are few or even none of labelled data in target domain, domain adaptation helps to reduce the amount of labelled data needed. Most traditional works [1–5] either learn a shared representation between source and target domain, or project features into a common subspace. Recently, there are also works [6–8] proposed to learn a scene-specific detector by deep architectures. However, these approaches are heuristic either on constructing feature space or re-weighting samples. Our motivation of developing a domain adaptation architecture is to reduce heuristic methods required during adaptation process.

In this paper, we proposed a new approach of unsupervised deep domain adaptation on people detection. Using source model trained on source domain as initialization, we utilize iterative algorithm to iteratively auto-annotate target examples with high confidence as people instance on target domain for the first iteration. During each iteration, these auto-annotated data are regarded as training set to update target model, which, then, can be taken as the auto-annotation tool to auto-annotate target samples for the next iteration. However, these auto-annotated samples are defective, including lack of negative samples and existence of false positive samples, which will no doubt lead to exploration of predictions on non-people instances. Therefore, on the one hand, to compensate for the quantitative imbalance between positive and negative samples, we randomly sample negative instances from source domain and mix into training set. On the other hand, based on deep network, we design an unsupervised regularizer to mitigate influence from data error and avoid overfitting. To have better regularization effect during adaptation process, we transform the last full connected layer of deep model into two sub-layers, element-wise multiply layer and sum layer. Thus, the unsupervised regularizer can be added on element-wise multiply layer to adjust all weights in the deep network and gain better performance.

Also, we further evaluate our approach on standard domain adaptation benchmark Office Dataset. The results of our adaptation approach outperform previously published works on both supervised and unsupervised scenarios, which also demonstrate the feasibility of our adaptation approach on both detection and classification tasks.

The contributions of our work are three folds.

- We proposed a feasible scheme to learn scene-specific deep detectors for target domains by unsupervised methodology, which can be easily deployed to various surveillance situations without any additional annotations.
- For better performance of unsupervised regularizer, we transform the last full connected layer of deep network into two sub-layers, element-wise layer and sum layer. Thus, all weights contained in the deep network can be adjusted under the unsupervised regularizer. To our knowledge, this is the first attempt to transform full connected layers for the purpose of domain adaptation.
- Experiments on standard domain adaptation benchmarks for classification also demonstrate the applicability of our approach to other deep domain adaptation tasks.

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 presents the details of our approach. Experimental results are shown in Section 4. Section 5 concludes the paper.

2 Relate Work

In many detection works, generic model trained by large amount of samples on source domain are directly utilized to detect on target domain. They assume that samples on target domain are subsets of source domain. However, when the distribution of data on target and source domain varies largely, the performance will drop significantly. Domain adaptation aims to reduce the amount of data needed for target domain.

Many domain adaptation works tried to learn a common representation space shared between source and target domain. Saenko et al. [1, 2] proposed a both linear-transform-based technique and kernel-transform-based technique to minimize domain changes. Gopalan et al. [3] projected features into Grassmann manifold instead of operating on features of raw data. Alternatively, Mesnil et al. [9] used transfer learning to obtain good representations. However, these methods are limited because scene-specific features are not learned to boost accuracy. The regularizer of our method are inspired these works.

Another group of works [4, 5, 10] on domain adaptation is to make the distribution of source and target domain more similar. Among these works, Maximum Mean Discrepancy (MMD) is used to as a metric to reselect samples from source domain in order to have similar distribution as target samples. In [11], MMD is incorporated as regularization to reduce distribution mismatch.

There are also works on deep adaptation to construct scene-specific detector. Wang et al.[6] explored context cues to compute confidence, and [7] learns distributions of target samples and proposed a cluster layer for scene-specific visual patterns. These works re-weighted auto-annotated samples for their final object function and additional context cues are needed for reliable performance. However, heuristic methods are required to re-weight samples. Alternately, Hattori el al. [8] learned scene-specific detector by generating a spatially-varying pedestrian appearance model. And Pishchulin et al. [12] used 3D shape models to generate training data. However, Synthesis for domain adaptation are also costly. Our approach minimize heuristic algorithms needed during adaptation process.

3 Our Approach

In this section, we introduce our unsupervised domain adaptation architecture on the task of people detection in crowded scenes. Under the unsupervised setting, we use iterative algorithm to iteratively auto-annotate target samples and update target model. As the auto-annotated samples contain data error, the performance decline caused by false positive samples will exceed the performance boost resulting from true positive samples if the target model is trained only

with detection loss function. Therefore, an unsupervised regularizer is required to mitigate the influence from data error on target model. Based on the assumption that source domain and target domain should share the same feature space after feature extraction layers, we encode an unsupervised regularizer to make a constraint that the distribution of data representation on the element-wise multiply layer should be similar between source stream and target stream.

The adaptation architecture of our approach consists of two streams – source stream and target stream, as shown in Fig 1. Source stream takes samples from source domain as input, while target stream are trained by auto-annotated positive samples from target domain and negative samples from source domain. These two streams can utilize any deep detection network as their basic model, as well as their detection loss function as supervised loss functions of two streams. In our experiment, we use the detection network mentioned in Section 3.3 as the basic model. For the purpose of unsupervised regularizer, the last full connected layers on both source and target model are transformed into two sub-layers – element-wise multiply layer and sum layer, as mentioned in Section 3.2.

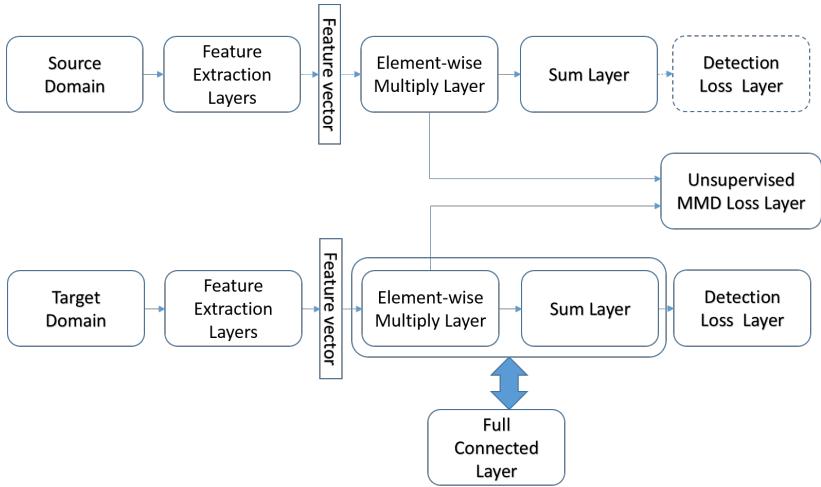


Fig. 1. xxxxxxxxxxxxxxxxxxxxxxxxxx

The loss function of adaptation architecture is composed of a supervised loss and an unsupervised regularizer. We denote training samples from source domain as $\mathbf{X}^S = \{x_i^S\}_{i=1}^{N^S}$. For training samples on source domain, we have corresponding annotations $Y^S = \{y_i^S\}_{i=1}^{N^S}$ with $y_i^S = (b_i^S, l_i^S)$, where $b_i^S = (x, y, w, h) \in R^4$ is the bounding box location and $l_i^S \in \{0, 1\}$ is the label indicating whether x_i^S is a people instance. At n^{th} adaptation iteration, we have two set of training samples, auto-annotated positive samples from target domain $\mathbf{X}^{T,n} = \{x_j^{T,n}\}_{j=1}^{N^{T,n}}$

and tantamount negative samples from source domain $\mathbf{X}^{S,n} = \{x_k^{S,n}\}_{k=1}^{N^{T,n}}$. Their corresponding annotations can be denoted as $Y^{T,n} = \{y_j^{T,n}\}_{j=1}^{N^{T,n}}$ and $Y^{S,n} = \{y_k^{S,n}\}_{k=1}^{N^{T,n}}$ with $y_j^{T,n} = (b_j^{T,n}, l_j^{T,n} \equiv 1, c_j^{T,n})$, and $y_k^{S,n} = (b_k^{S,n}, l_k^{S,n} \equiv 0)$, respectively. $c_*^{T,n}$ is the confidence given by auto-annotation tools and N^I is the maximum number of adaptation iterations. Now we can formulate the combination of supervised loss and unsupervised regularizer as follows:

$$L(\theta^{T,n} | \mathbf{X}^{T,n}, \mathbf{Y}^{T,n}, \mathbf{X}^{S,n}, \mathbf{Y}^{S,n}, \mathbf{X}^S, \theta^S) = L_S + \alpha * L_U \quad (1)$$

$$\begin{aligned} L_S &= \sum_{j=1}^{N^{T,n}} \sigma(c_j^{T,n}) * (R(\theta^{T,n} | x_j^{T,n}, b_j^{T,n}) + C(\theta^{T,n} | x_j^{T,n}, l_j^{T,n})) \\ &\quad + \sum_{k=1}^{N^{T,n}} (R(\theta^{T,n} | x_k^{S,n}, b_k^{S,n}) + C(\theta^{T,n} | x_k^{S,n}, l_k^{S,n})) \end{aligned} \quad (2)$$

$$L_U = L_{EW}(\theta^T | \mathbf{X}^T, \mathbf{X}^S, \theta^S) \quad (3)$$

where L_S is supervised loss to learn the scene-specific detector and L_U is the unsupervised regularizer part. $\sigma(\cdot)$ is a step function to re-weight supervised loss of auto-annotated data. $R(\cdot)$ is a regression loss for bounding box location, such as norm-1 loss, and $C(\cdot)$ is a classification loss for bounding box confidence, such as cross-entropy loss. And $L_{EW}(\cdot)$, to be introduced in Section 3.2, is the MMD-based loss added on the element-wise multiply layer for unsupervised regularization. α is the coefficient balancing the effect of supervised and unsupervised loss.

3.1 Iterative Algorithm

Iterative algorithm is the training method of the target stream of our adaptation architecture. There are two reasons to employ iterative algorithm. Firstly, auto-annotated data on target domain vary for every adaptation iteration and new positive samples will be auto-annotated as training set. Compared to methods without iterative algorithm, it helps to avoid overfitting caused by lack of data. Besides, unsupervised regularizer performs better with more training data as it's a distribution based regularizer.

There are two stages for iterative algorithm. Source stream and target stream are separately trained at different stages. At initialization stage, source model of source stream are trained under supervised loss function with abundant labelled data, \mathbf{X}^S , from source domain. After its convergence, the weights of source model θ^S are taken to initialize target stream. At adaptation stage, target model is trained by auto-annotated data under both supervised loss function and unsupervised regularizer. Note that we do not jointly train two streams at adaptation stage and the weights of source model stays static which is served as a distribution reference for unsupervised regularizer at adaptation stage. The complete adaptation process is illustrated in Algorithm 1. After a predetermined iteration limit N^I is reached, we obtain our final detection model on the target domain.

Algorithm 1 Deep domain adaptation algorithm (to be completed)

```

225 1: procedure DEEP DOMAIN ADAPTATION
226 2: Train source model on source stream with abundant annotated data
227 3: Use well-trained source model on source stream to initialize target model on target
228   stream as  $M_0$ 
229 4:   for  $i = 0:N^I$  do
230 5:      $M_i$  generate auto-annotated positive samples  $\mathbf{X}^{T,n}$  of target domain with  $\mathbf{Y}^{T,n}$ 
231 6:     Randomly sampled negative instances  $\mathbf{X}^{S,n}$  from source domain with  $\mathbf{Y}^{S,n}$ 
232 7:      $\mathbf{X}^n = \{\mathbf{X}^{T,n}, \mathbf{X}^{S,n}\}$ 
233 8:      $\mathbf{Y}^n = \{\mathbf{Y}^{T,n}, \mathbf{Y}^{S,n}\}$ 
234 9:     Take  $(\mathbf{X}^n, \mathbf{Y}^n)$  as training data to upgrade  $M_i$  into  $M_{i+1}$ 
235 10:   end for
236 11:  $M_{NI}$ : final model.
237 12: end procedure
238
239
```

Different from source domain, we use auto-annotation tools to auto-annotate people instances with high confidence as training set on target domain. At first iteration, we take source model as the auto-annotation tool. For subsequent adaptation iterations, updated target model from last adaptation iteration are utilized as the auto-annotation tool. During every adaptation iteration, the auto-annotation tool (also the target model) will be updated. Thus, the training samples for target domain at n^{th} adaptation iteration may differ from that at $(n + 1)^{th}$ adaptation iteration.

Note that these auto-annotated data are all regarded as positive samples. That is, we encourage the network to have more confidence on positive instances and stay conservative toward negative instances. This policy will no doubt resulting predictions of many non-people instance and impeding the further training when many non-people instances are regarded as people instances by the target model. To compensate for lack of negative instances, equal amount of negative samples are randomly selected from target domain, which are human annotated and can thus provide true negative samples. To further mitigate influence from data error in auto-annotated data, we will introduce an unsupervised regularizer in Section 3.2.

259 3.2 Unsupervised weights regularizer on Element-wise Multiply 260 Layer

Element-wise Multiply Layer In deep neural network, the last feature vector layer are taken as an important data representation of input images. However, in this paper, we take one step further to focus on the last full connected layer which serves as an decoder to decode rich information contained in the last feature vector into final outputs. As source model are trained with abundant labelled data on source domain, the parameters of the last full connected layer are also well converged. We assume that a regularizer on the last full connected layer will achieve better results compared with the last feature vector layer.

270 Firstly, denote the last feature vector, parameters of the last full connected layer
 271 and final outputs as $\mathbf{F}_{(N^B \times N^D)}$, $\mathbf{C}_{(N^D \times N^O)}$ and $\mathbf{P}_{(N^B \times N^O)}$, respectively. The
 272 operation of full connected layer can be thus formulated as matrix multiply:
 273

$$\mathbf{P} = \mathbf{F} * \mathbf{C} \quad (4)$$

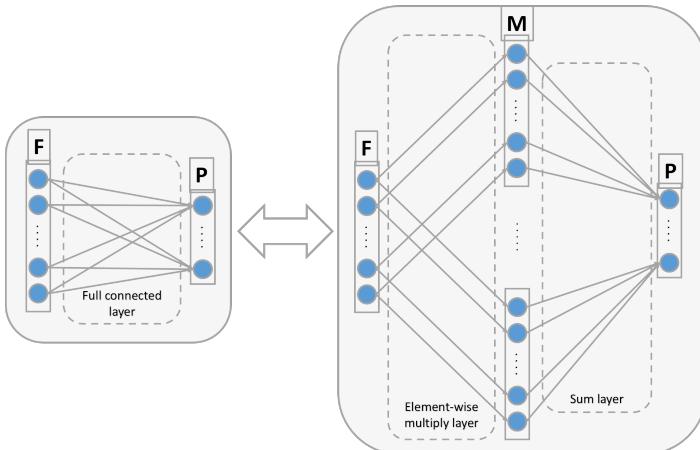
$$P_{b,o} = \sum_d F_{b,d} * K_{d,o} \quad (5)$$

277 Inspired by this form, we separate the above formula into two sub-operations –
 278 element-wise multiply and sum, which can be formulated as:
 279

$$M_{b,o,d} = F_{b,d} * C_{d,o} \quad (6)$$

$$P_{b,o} = \sum_d M_{b,o,d} \quad (7)$$

284 where $\mathbf{M}_{(N^B \times N^O \times N^D)} = [\mathbf{m}_{b,o}]$ is the parameter tensor of element-wise multi-
 285 ply operations. $\mathbf{m}_{b,o}$ is a vector with N^D dimensions, which will be the basis
 286 of unsupervised regularizer. Finally, we can equivalent-transform the last ful-
 287 l connected layer between the last feature vector layer and final outputs layer
 288 into element-wise multiply layer and sum layer. The transformed element-wise
 289 multiply layer is thus the last layer with parameters before output layers. Fig 2
 290 illustrates the transform.
 291



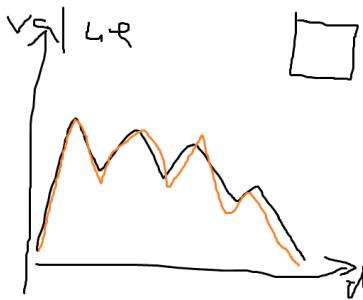
308 **Fig. 2.** xxxxxxxxxxxxxxxxx

312 **Unsupervised weights regularizer on Element-wise Multiply Layer** In
 313 works [decaf][1], the last feature vector layer are regarded as the final representa-
 314 tion of images. (how to introduce beyond sharing weights) In domain adaptation

315 tasks, when generic deep model are trained with abundant data from source do-
 316 main, the last element-wise multiply layer mentioned in Sec 3.2 also includes
 317 rich information leading to the final output. For the nodes in the output layer,
 318 inputs from element-wise multiply layers that will contribute to its value are
 319 not randomly decided. In this paper, we assume that the distribution of $\mathbf{m}_{b,o}$
 320 of the last element-wise multiply layer on both source and target domain should
 321 be similar. In the last element-wise multiply layer, every dimension of $\mathbf{m}_{b,o}$ may
 322 contribute to the final output. However, as the old model on source domain
 323 are trained with abundant images, the distribution of $\mathbf{m}_{b,o}^T$ should be consistent
 324 compared with $\mathbf{m}_{b,o}^S$ of source domain when adapting deep network on target
 325 domain. We utilize MMD (maximum mean discrepancy) to encode the similarity
 326 of element-wise multiply layer of source domain and target domain:

$$327 \quad L_{MMD}(\theta^T | \mathbf{X}^S, \mathbf{X}^T, \theta^S) = \frac{1}{N^O} \sum_{o=1}^{N^O} \left\| \frac{1}{N^B} \sum_{b=1}^{N^B} \mathbf{m}_{b,o}^T - \frac{1}{N^B} \sum_{b=1}^{N^B} \mathbf{m}_{b,o}^S \right\|^2 \quad (8)$$

330 which can also interpreted as the Euclidean distance between the center of $\mathbf{m}_{b,o}^T$
 331 and $\mathbf{m}_{b,o}^S$ across all output nodes. It's unpractical to get the distribution of
 332 the whole training set, while too few images cannot obtain a stable center for
 333 regularization. In our experiments, the $L_{MMD}(\cdot)$ loss is calculated for every
 334 batch. An example comparison of centers of $\mathbf{m}_{b_i,o}^S$ and $\mathbf{m}_{b_j,o}^S$ are shown in Fig
 335 x(1).



350 **Fig. 3.** One kernel at x_s (dotted kernel) or two kernels at x_i and x_j (left and right)
 351 lead to the same summed estimate at x_s .

355 3.3 Detection Network

356 The generic model ¹ used in our adaptation architecture for source and tar-
 357 get stream is an end to end detection network without any precomputed region

358 ¹ Proposed by Russel et al.

proposals needed. It consists of a GoogLeNet [13] for feature extraction and a RNN-based decoder for output of bounding box and corresponding confidence, as shown in Fig x(!). Firstly, the GoogLeNet model encode the image into a feature map ($15 \times 20 \times 1024$) of which each 1024 dimension vector are data representation of its receptive field corresponding to a subregion of the image. Then the RNN-based layers with batch size 300 will decode the data representation and sequentially predict 5 possible bounding boxes by the order of its corresponding confidence. Finally, all outputs are summarized to give final detection results. When trained with abundant samples from source domain, the obtained model has a high precision on target domain, however, its recall is low. Also, different from other detection network [14, 15], which need precomputed proposals for classification and fine regression, this generic model directly predict bounding boxes with high confidence. Thus, negative instances may contain people and non-people predictions, and cannot be employed in adaptation training.

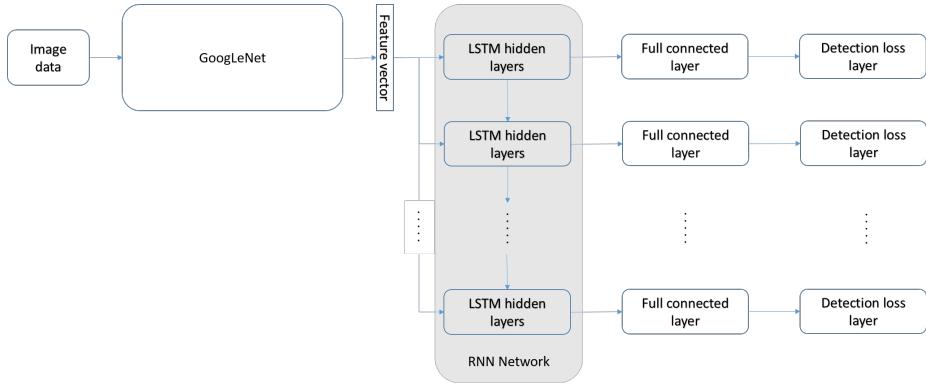


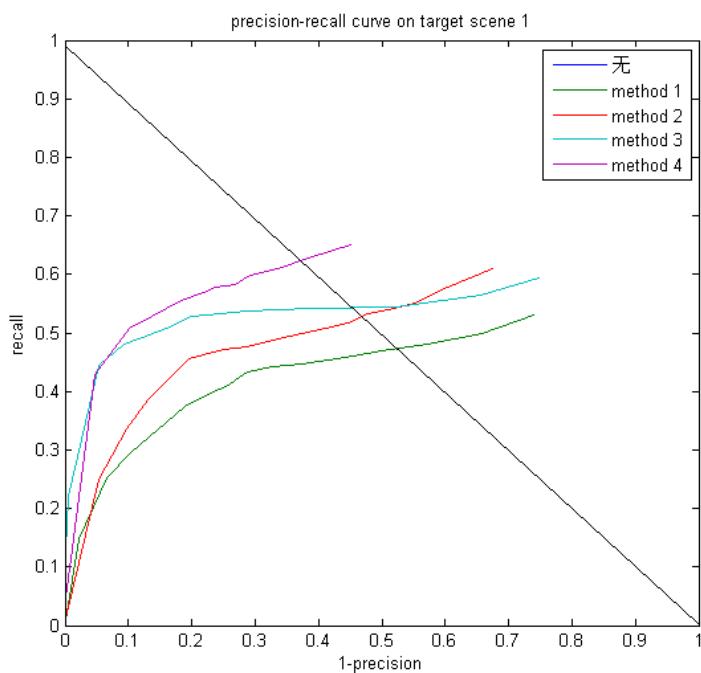
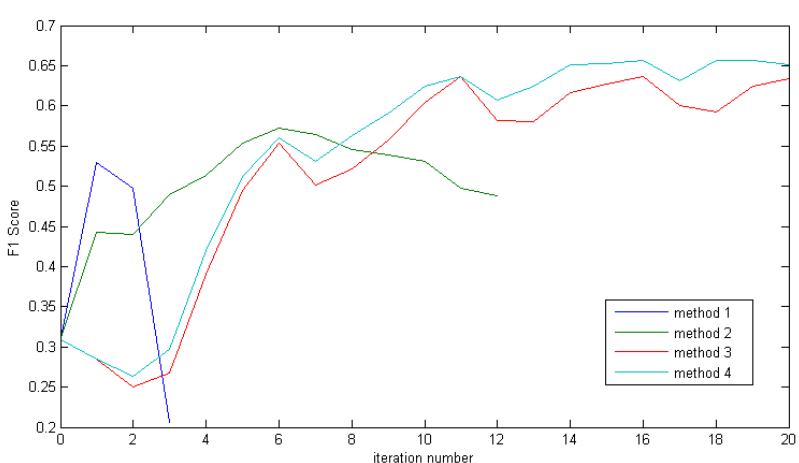
Fig. 4. xxxxxxxxxxxxxxxx

4 Experiment Results

In this section, we introduce experiment results on both surveillance applications and standard domain adaptation dataset. Our motivation for unsupervised domain adaptation method is for easier deployment of We firstly evaluate our approach on video surveillance. Then we employ our approach to standard domain adaptation benchmarks on both supervised and unsupervised settings to demonstrate the effectiveness of our method.

4.1 Domain Adaptation on Crowd Dataset

Dataset and evaluation metrics To show the effectiveness of our domain adaptation approach on people detection, we collected a dataset consisting of

429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
Fig. 5. xxxx448
449
Fig. 6. xxxx

450 3 target scenes for target domain. These three scenes contain 1308, 1213 and
451 331 un-annotated images with 0000, 0000, 0000 people instances respectively.
452 For each scene, 100 images are annotated for evaluation. Instead of labelling the
453 whole body of a person, we labels the head of a person as bounding box during
454 training. The motivation for labelling only people heads comes from detection of
455 indoor people or in crowded scenes, where the body of a person may be invisible.
456 The dataset for source domain are Brainwash Dataset released on [http://d2.mpi-](http://d2.mpi-inf.mpg.de/datasets)
457 inf.mpg.de/datasets. Brainwash Dataset consists of over 11917 images from 3
458 crowded scenes. Examples of images from source and target domain are shown
459 in Fig x(!).

460 Our evaluation metrics for detection uses the protocol defined in PASCAL
461 VOC [16]. To judge a predicted bounding box whether correctly matches a
462 ground truth bounding box, their intersection over their union must exceed 50%.
463 And Multiple detections of the same ground truth bounding box are regarded as
464 one correct prediction. We plot the precision-recall curve in Fig x(!). Also, the
465 F1 score $F1 = 2 * precision * recall / (precision + recall)$ during the adaptation
466 process are also shown in Fig x(!).

467
468 **Experimental settings** We use deep learning framework Caffe [17] as the
469 adaptation architecture of our approach. During the adaptation, we set learning
470 rate as 0.01 and momentum as 0.5. At initialization stage, GoogLeNet weights
471 are firstly used to initialize source model of source stream, while parameters in
472 RNN layers are randomly initialized from a uniform distribution. For each iteration,
473 100 auto-annotated images from target domain and 1000 annotated images
474 from source domain are alternatively used for training. The outputs of our
475 detection network contains bounding box locations and corresponding confidence,
476 thus there are two full connected layers between the last feature vector layer
477 and the final outputs. Experiments of unsupervised regularizer on element-wise
478 multiply layer for bounding box regression have no additional improvement on
479 the performance when regularizer on element-wise multiply layer for box
480 confidence classification are added already. Our approach on domain adaptation are
481 executed separately in the 3 target scenes.

482
483 **Comparison with baseline methods** To demonstrate the effectiveness of our
484 approach, 4 methods are compared with method 4 as our final approach:
485
486

- 487 1. Only auto-labeled samples on target domain are used for training, and with-
488 out any unsupervised regularizer.
- 489 2. Only auto-labeled samples on target domain are used for training, with M-
490 MD regularizer on last element-wise multiply layer as unsupervised weights
491 regularizer.
- 492 3. Both auto-labeled images from target domain and labeled images from source
493 domain are alternately sampled for training, with MMD regularizer on last
494 feature vector as unsupervised weights regularizer.

- 495 4. Both auto-labeled images from target domain and labeled images from source
 496 domain are alternately sampled for training, with MMD regularizer on last
 497 element-wise multiply layer as unsupervised weights regularizer.

498 Fig x(!) plots the precision-recall curve of the above comparison methods in
 499 target scene 1). Also, the F1 score changes of every iteration during adaptation
 500 process are also depicted in Fig x(!). Table x(!) gives concrete precision and recall
 501 value of the 4 comparison methods on three target scenes when the F1 scores
 502 are at their highest. Examples of adaptation results are shown in Fig x(!).

	Scene 1			Scene 2			Scene 3		
	1-Pr	Re	F1	1-Pr	Re	F1	1-Pr	Re	F1
method 0	0.101	0.187	0.309	0.015	0.683	0.807	0.035	0.412	0.577
method 1	0.245	0.408	0.530	0.632	0.905	0.524	0.176	0.778	0.800
method 2	0.284	0.476	0.572	0.012	0.837	0.906	0.078	0.653	0.764
method 3	0.109	0.496	0.637	0.002	0.721	0.838	0.044	0.611	0.746
method 4	0.140	0.530	0.656	0.006	0.811	0.893	0.097	0.778	0.836

512
 513
 514
 515 **Performance evaluation** From the Table x(!) and Fig x(1), we have the fol-
 516 lowing observations:

- 517 – The recall values of method 1,2,3,4, which all utilized iterative algorithm
 518 to upgrade the target model, are larger than that of method 0, which are
 519 source model trained on source domain. This implies the effectiveness of our
 520 iterative algorithm in auto-annotation and iterative training.
 521 – Compared with method 1, method 2 has higher F1 score. Their difference
 522 on whether a MMD regularizer are added into loss function demonstrates
 523 that our unsupervised regularizer can suppress data error and thus boost
 524 the final recall.
 525 – Method 4 has both higher precision and higher recall than method 2, which
 526 demonstrates the effectiveness of additional samples from source domain
 527 during adaptation process.
 528 – Compared with method 3, the recall of method 4 are further boosted. This
 529 results from the transformed element-wise multiply layer which provided
 530 better regularization effect on the target model.

532 4.2 Domain Adaptation on Standard Classification Benchmark

533
 534 **Office dataset** The Office dataset [1] comprises 31 categories of objects from
 535 3 domains (Amazon, DSLR, Webcam). Example images are depicted in Fig.
 536 x(!). As Amazon domain contains 2817 labelled images, which is the largest, we
 537 take it as source domain and Webcam domain as target domain. We follow the
 538 standard protocol for both supervised and unsupervised settings. Specifically,
 539 for supervised domain adaptation, we use 20 randomly sampled images with

540 labels for each category as training data for Amazon domain. When evaluate
 541 on unsupervised domain adaptation, 3 labelled images from target domain are
 542 additional selected for each class. For both settings, the rest of images on target
 543 domain are used for evaluation.



544
 545
 546
 547
Fig. 6. Some examples from three domains in the Office dataset.
 548

549
 550
 551
Fig. 7. One kernel at x_s (*dotted kernel*) or two kernels at x_i and x_j (*left and right*)
 552

553
 554
 555
 556 **Experimental settings and network design** On supervised setting, we
 557 reused the architecture in people detection. We utilize AlexNet [18] as the generic
 558 model of both streams. Firstly, we train the source model on source stream
 559 with provided training data from Amazon domain. Then iterative algorithm
 560 mentioned in Sec 3.1 are utilized for adaptation. The auto-labelling tool takes
 561 images on target domain with confidence exceeding 0.9 as training data for next
 562 iteration. The difference is besides auto-labelled images on target domain, 3 hu-
 563 man labelled images for each class are also included as training data for target
 564 model. For each iteration, 100 images are randomly sampled from training data.
 565 The unsupervised MMD regularizer added on the element-wise multiply layer
 566 transformed from the last full connect layer of target model set the coefficient
 567 value α as 10 on Eq 3, the same as that in people detection task.
 568

569 We use the same experimental setting for our unsupervised adaptation, ex-
 570 cept that at adaptation stage, no human labelled images can be added into the
 571 training set.

572
 573
 574
 575
 576 **Performance evaluation** In Table x(!), we compare our approach with oth-
 577 er six recently published works in both supervised and unsupervised settings.
 578 The outstanding performance on both settings confirms the effectiveness of our
 579 iterative algorithm and MMD regularizer on the element-wise multiply layer
 580 transformed from the last full connect layer.

581 5 Conclusions

582
 583 The paper ends with a conclusion.

	$A \rightarrow W$	
	Supervised	Unsupervised
GFK(PLS,PCA)[19]	46.4	15.0
SA [20]	45.0	15.3
DA-NBNN [21]	52.8	23.3
DLID [22]	51.9	26.1
DeCAF ₆ S [23]	80.7	52.2
DaNN [11]	53.6	35.0
Ours	84.3	66.3
Ours	85.4	69.3



Fig. 8. One kernel at x_s (*dotted kernel*) or two kernels at x_i and x_j (*left and right*)

630 References

- 631 1. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to
new domains. In: Computer Vision–ECCV 2010. Springer (2010) 213–226
- 632 2. Kulis, B., Saenko, K., Darrell, T.: What you saw is not what you get: Domain
adaptation using asymmetric kernel transforms. In: Computer Vision and Pattern
Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1785–1792
- 633 3. Gopalan, R., Li, R., Chellappa, R.: Domain adaptation for object recognition: An
unsupervised approach. In: Computer Vision (ICCV), 2011 IEEE International
Conference on, IEEE (2011) 999–1006
- 634 4. Huang, J., Gretton, A., Borgwardt, K.M., Schölkopf, B., Smola, A.J.: Correcting
sample selection bias by unlabeled data. In: Advances in neural information
processing systems. (2006) 601–608
- 635 5. Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., Schölkopf, B.:
Covariate shift by kernel mean matching. Dataset shift in machine learning **3**(4)
(2009) 5
- 636 6. Wang, X., Wang, M., Li, W.: Scene-specific pedestrian detection for static video
surveillance. Pattern Analysis and Machine Intelligence, IEEE Transactions on
36(2) (2014) 361–374
- 637 7. Zeng, X., Ouyang, W., Wang, M., Wang, X.: Deep learning of scene-specific clas-
sifier for pedestrian detection. In: Computer Vision–ECCV 2014. Springer (2014)
472–487
- 638 8. Hattori, H., Naresh Boddeti, V., Kitani, K.M., Kanade, T.: Learning scene-specific
pedestrian detectors without real data. In: Proceedings of the IEEE Conference
on Computer Vision and Pattern Recognition. (2015) 3819–3827
- 639 9. Mesnil, G., Dauphin, Y., Glorot, X., Rifai, S., Bengio, Y., Goodfellow, I.J., Lavoie,
E., Muller, X., Desjardins, G., Warde-Farley, D., et al.: Unsupervised and transfer
learning challenge: a deep learning approach. ICML Unsupervised and Transfer
Learning **27** (2012) 97–110
- 640 10. Gong, B., Grauman, K., Sha, F.: Connecting the dots with landmarks: Discrimin-
atively learning domain-invariant features for unsupervised domain adaptation.
In: Proceedings of The 30th International Conference on Machine Learning. (2013)
222–230
- 641 11. Ghifary, M., Kleijn, W.B., Zhang, M.: Domain adaptive neural networks for object
recognition. In: PRICAI 2014: Trends in Artificial Intelligence. Springer (2014)
898–904
- 642 12. Pishchulin, L., Jain, A., Wojek, C., Andriluka, M., Thormählen, T., Schiele, B.:
Learning people detection models from few training samples. In: Computer Vision
and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1473–
1480
- 643 13. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D.,
Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings
of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1–9
- 644 14. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on
Computer Vision. (2015) 1440–1448
- 645 15. Vu, T.H., Osokin, A., Laptev, I.: Context-aware cnns for person head detection.
In: Proceedings of the IEEE International Conference on Computer Vision. (2015)
2893–2901
- 646 16. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisser-
man, A.: The pascal visual object classes challenge: A retrospective. International
Journal of Computer Vision **111**(1) (2015) 98–136

17. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
 18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
 19. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2066–2073
 20. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 2960–2967
 21. Tommasi, T., Caputo, B.: Frustratingly easy nbnn domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 897–904
 22. Chopra, S., Balakrishnan, S., Gopalan, R.: Dlid: Deep learning for domain adaptation by interpolating between domains. In: ICML workshop on challenges in representation learning. Volume 2. (2013)
 23. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531 (2013)