Abstract

The current advances in object detection depend on large-scale datasets to get good performance. However, there may not always be sufficient samples in many scenarios, which leads to the research on few-shot detection as well as its extreme variation one-shot detection. In this paper, the one-shot detection has been formulated as a conditional probability problem. With this insight, a novel one-shot conditional object detection (OSCD) framework, referred as Comparison Network (ComparisonNet), has been proposed. Specifically, query and target image features are extracted through a Siamese network as mapped metrics of marginal probabilities. A two-stage detector for OSCD is introduced to compare the extracted query and target features with the learnable metric to approach the optimized non-linear conditional probability. Once trained, ComparisonNet can detect objects of both seen and unseen classes without further training, which also has the advantages including class-agnostic, training-free for unseen classes, and without catastrophic forgetting. Experiments show that the proposed approach achieves state-of-the-art performance on the proposed datasets of Fashion-MNIST and PASCAL VOC.

1. Introduction

Deep learning based models have achieved great advances in object detection [1] in recent years. However, these methods have to rely on large amounts of labelled data for each object category during the training period, which have the inherent shortcomings in practical applications. For instance, collecting a large number of annotated samples is quite time-consuming and laborious. And in the real-world scenarios, it could be challenging to obtain enough annotated samples for some specific object categories (e.g., rare animals). Under this situation, models suffer from severe overfitting due to limited samples. Moreover, these fully supervised learning models are hard to extend to new classes. If we want to add a new object category for detection, we have to retrain the entire model or at least the subnetwork for detection. Due to these inherent properties, the current object detection methods are intractable for the task of detecting new object classes with...
few samples.

In contrast, humans are good at learning new concepts from one or a few instances. For example, a person can easily identify an aeroplane after given a single picture. Inspired by the ability of humans in one- and few-shot learning, researches on few-shot image classification and detection start to emerge \[9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21\]. While the same time, few work \[22, 23\] focus on the one-shot image detection, i.e., an extreme scenario of few-shot learning, which is useful in some applications with only one input samples from the newly interested class, e.g., the mentioned rare animal detection. Similarly, in this paper, we propose a novel one-shot conditional detection framework with generalization capability, which is hinted by the conditional probability estimation theory.

As shown in Figure 1 (a), the traditional framework of one-shot detection proposed in \[22, 23\] treats the one-shot detection problem as a classical detection problem with sliding window scheme and chooses the best fitted window. Especially a metric learning module, instead of the classical classifiers, is utilized to calculate the similarity between the query image from the newly interested class and the tagged window of testing sample. As another approach shown in Figure 1 (b), the LSTD \[15\] framework chooses to utilize samples from the newly interested class to finetune both the region proposal network (RPN, i.e., to replace traditional sliding window scheme) and deep learning based classifier with the expectation that the original detection flow can adaptively recognize new classes with limited number of finetune samples.

Hinted by the essential similarity between these two approaches, it is revealed that both approaches can be equalized with a conditional probability estimation model, by treating the original detection scheme as the independent probability estimation and the query image from new class as condition. Then, under the guidance of the famous Bayes conditional probability theory, this paper proposes that the one-shot detection problem should be better named as One-Shot Conditional Detection (OSCD) (Figure 2) and designs a generalization OSDC methodology based on comparison learning \[14\], shown in Figure 1 (c) and was named as ComparisonNet in this paper. In the proposed ComparisonNet, a Siamese subnetwork is first employed to extract the query and target features as marginal probabilities. Then the carefully designed conditional region proposal network (C-RPN) and learnable metric will work in a two-stage pipeline to estimate the conditional probability. Such a framework can understand as a meta learning process to approach the optimized non-linear conditional probability estimation method.

The proposed methodology is theory guided and works from coarse to fine to detect objects beyond the seen classes with a single example in an unsupervised manner. The advantages of class-agnostic, training-free, and overcoming forgetting can be expected: Class-agnostic: based on metric learning, the OSDC model only computes the similarity score between the query and target images. It doesn’t care which category the query image is. Training-free: once trained, ComparisonNet can detect unseen class objects without further training on the samples of unseen classes. Overcome forgetting: the proposed model never forget the knowledge of training dataset due to no further updating during test. Experiments on two OSCD benchmarks prove the effectiveness of our method with significantly increased accuracy.

The contributions of this paper are summarized as follows:

1. A conditional probability estimation based theory to model the OSCD problem, which can guide the implementation of algorithm design to approach the optimized performance.
2. A unified, end-to-end framework named ComparisonNet for OSCD, which is class-agnostic, training-free for unseen classes, and never forget the knowledge of seen classes.
3. Two state-of-the-art OSCD datasets with significantly increased number of classes and image sizes, which are proposed to improve the test performance for the community.

2. Related Works

Since this work involves object detection and few-shot learning, well introduce related studies in these areas.

2.1. Object Detection

Convolutional Neural Network (CNN) based object detection approaches can mainly be divided into two categories: one-stage detectors \[1, 4, 3, 5\] and two-stage detectors \[2, 7, 6, 8\]. One-stage detectors predict object locations and classes in the whole image directly, without region proposal procedure. While two-stage detectors first use region proposal methods to generate a set of candidate object locations, then classify each candidate location as foreground classes or background.

2.2. Few-Shot Learning

One- or few-shot learning methods aim to learn new knowledge rapidly from few data. Considering that deep models \[5, 8, 24, 25, 26, 27\] trained on the data-rich datasets have led to significant advances and universal applications in image classification and object detection, fine-tuning pretrained models can be a simple and efficient method to transfer knowledge from source domains to target domains. Beyond the basic fine-tuning operation, metric learning based works \[13, 20, 21\] and meta-learning based methods \[9, 10, 11, 12, 14, 16, 17, 18, 19\] have made great
progress in few-shot learning. Our method is inspired by Relation Network [14]. Unlike most previous methods using pre-defined distance metrics, Relation Network proves that learnable distance metric outperforms pre-defined distance metrics in few-shot learning.

Compared to the advances in few-shot image classification, few works have addressed on few-shot object detection so far. [28] uses few-shot detection in a semi-supervised learning framework. The transfer learning method LSTD [15] aims to detect objects with few training samples per category.

2.3. One-Shot Conditional Object Detection

One-shot learning is a basic task and it can be easily extended to few-shot learning [14]. Performance on few-shot detection will improve once there are advances in OSCD. Therefore, we focus on OSCD in this paper. OSCD is consistent with some visual searching tasks such as template matching [29, 30, 31], image retrieval [32, 33, 34, 35] and person re-identification [36, 37, 38], in principle. While it is implemented with an object detection pipeline in the same manner as the common object detection tasks.

Previous works [22, 23] for OSCD follow the sliding-window paradigm, in which a classifier is applied on a dense image grid. They use handcrafted features to represent image patches. If a grid and the query image have a high similarity in feature space, the grid may contain an object instance. Whereas the sliding-window strategy is not flexible for different scale and aspect ratio objects. Furthermore, hand-engineered features tend to have poor performance for viewpoint and intra-class variation.

2.4. Visual Tracking

We note that the task of video object tracking [39, 40] is similar to one-shot conditional object detection to some extent (see section 4.2 in [39]). However, there are two main differences between the two tasks. Firstly, in video tracking, we only search for the same object in the next frame. But for OSCD, we search for all objects belonging to the query class. Intra-class variation requires OSCD to learn a more general representation ability. Secondly, the prior knowledge of object location in the last frame helps to narrow the searching scope to a smaller neighborhood, while one-shot conditional object detection has to conduct global searching within the whole image. Therefore, video tracking can be seen as a simplified local OSCD problem. These challenges make OSCD more complex than video tracking in principle.

2.5. Fixed Metric

It should be noted that both the classical OSCD work [22, 23], and the mentioned visual tracking work [40, 39] implement the similarity calculation by comparing the target features \(f_t\) and query features \(q_t\) with a pre-defined metric \(M(f_t, f_q)\). [22, 23] detect objects by computing the cosine similarity between the query and target features. The correlation operation in [40, 39] can also be viewed as a fixed manually defined metric. Although these metric based detection methods have made significant progress, the fixed defined metric may not be the best choice in the few-shot regime [14]. Besides, discovering an appropriate metric \(M(f_t, f_q)\) for a specific task is laborious.

3. Method

3.1. Problem Formulation

For object detection, suppose that there are no sufficient samples in the interested classes, leading to poor performance for the common supervised learning methods. Besides, we may not know which categories exist in future tasks. The more serious challenge is that the "object" can be any patterns of interest. All these issues make the object detection task extremely hard for traditional approaches. Thus, OSCD is proposed to address the above problems.

The goal of OSCD is to detect the objects of both seen and unseen classes according to the given condition (a single query image). In the OSCD regime, a model is trained on many query-target image pairs of seen classes to acquire strong priors. Once trained, the model should output the regions which is similar to the query image.

3.2. One-Shot Conditional Object Detection

The state-of-the-art deep learning based object detection models need to be trained for hundreds of thousands iterations on large-scale datasets, which requires that we must know the category of test objects in advance. There are many weakly supervised object detection methods [41, 42, 43, 44, 45, 46, 47], that aim to alleviate the heavy burdens of data annotation. While the transfer-learning based object detection method [15] can reduce the data requirements on target domains by acquiring strong priors on source domains. We shown a common two-stage detector (Faster R-
Figure 3. The frameworks of a common object detector (Faster R-CNN) and the proposed OSCD method (ComparisonNet). Faster R-CNN can locate and recognize the objects of seen classes (person), but fails to classify the objects of unseen classes (horse). In contrast, the proposed ComparisonNet implements conditional probability estimation by C-RPN and C-Classifier to pay more attention to the objects of query class (horse) and filters out the irrelevant objects of other classes (person), given the condition (query image). The ⊕ symbol denotes selecting features φ(·) for each candidate region generated from RPN or C-RPN.

CNN) in Figure 3(a) and formulate all of the above methods as:

\[ P(p_i, bbox_i) = F(\phi(I)) \] (1)

Here \( \phi \) is a CNN based feature extractor, \( F \) is a one- or two-stage detector, which can output the probability \( p_i \) of each category for every predicted bounding box \( bbox_i \), given the test sample \( I \).

Differently, the proposed method has a novel definition termed as conditional object detection (Figure 3(b)), which can be formulated as:

\[ P(p_i, bbox_i | I_q) = F(\phi(I_t) | \phi(I_q)) \] (2)

Unlike common object detection methods, the conditional object detection model predict the location \( bbox_i \) and similarity score \( p_i \) of each object according to the given condition — the query image \( I_q \), in the target image \( I_t \).

In fact, conditional object detection is class-agnostic and training-free for unseen classes, this means the proposed model is more flexible than the common methods. Note that even the transfer-learning based low-shot detectors like LSTD [15] still need to know the test classes and further update the network during test. Benefit from the metric based framework, the proposed OSCD method can overcome the catastrophic forgetting (degrading performance on source domains), which usually occurs when the model are trained on target domains. Besides, the conditional object detection method can detect any patterns of interest in principle, which should be explored in the future.

3.3. Conditional Probability Estimation for OSCD

The solution of OSCD task can usually be generalized as Bayesian probability estimation and formulated as:

\[ P(I_t | I_q) = \frac{P(I_t, I_q)}{P(I_q)} \] (3)

Simply apply above equation to solve the OSCD problem is intractable due to the commonly existed strong non-linearity in vision objects. Therefore, we introduce a deep learning based conditional probability estimation function \( F_B \) to approach the optimized non-linear conditional probability estimation and reformulate it as:

\[ P(I_t | I_q) = F_B(\phi(I_t), \phi(I_q)) \] (4)

First, the query and target image \( I_q, I_t \) are mapped into an embedding space by \( \phi \), resulting in the marginal probability \( \phi(I_t) \) and \( \phi(I_q) \). Then, we employ a two-stage pipeline of C-RPN and conditional classifier (C-Classifier) to further
alleviate the difficulties of OSCD. Hence, $F_B$ can be computed in two steps as follows:

$$P(I_i|I_q) = F_{CR}(\phi(I_i), \phi(I_q))$$
$$P^i(I_i|I_q) = F_{CC}(\phi^i(I_i), \phi(I_q))$$

(5)

Here, $F_{CR}$, $F_{CC}$ represents C-RPN and C-Classifier respectively, $i$ denotes the $i-$th similar region generated from $F_{CR}$. We select $\phi^i(I_i)$ according to the output of $F_{CR}$.

**Learnable metric:** One of the key issues in this process is to find the potential regions in the testing images for similarity calculation. Some traditional region proposal methods, such as sliding window, Selective Search [48], and EdgeBoxes [49], require high computation complexity and produce a significantly large number of regions. The advanced RPN also produces many irrelevant regions with the query image. In contrast, the proposed C-RPN is a query dependent region generator. Both $F_{CR}$ and $F_{CC}$ are based on the learnable metric $\mathcal{M}(f_t, f_q, \theta)$ to gain the conditional probability estimation.

A learnable metric $\mathcal{M}(f_t, f_q, \theta)$ can adjust its parameters $\theta$ adaptively in a meta-learning manner from the query-target pairs [14]. The relation between $\phi(I_q)$ and $\phi(I_t)$ is consistent among many sample pairs. We employ a learnable metric to learn this relation and extend it to the unseen sample pairs. Once trained on a series of sample pairs, a learned metric can lead to the similar test performance as in the training set.

**Feature fusion:** In the proposed method, there are no strict restrictions on the spatial size of the query and target images to make the model more flexible. Therefore, executing conditional probability estimation with different spatial sizes is a challenge. We propose two feature fusion modules in $F_{CR}$ and $F_{CC}$ respectively for effectively calculating similarity in each spatial position.

### 3.4. Implementation

The proposed conditional probability estimation for OSCD is implemented with ComparisonNet as illustrated in Figure 3(b). We employ a Siamese network as $\phi$ to obtain the query and target features $\phi(I_q)$ and $\phi(I_t)$, which are regarded as the marginal probability by the following two-stage pipeline of C-RPN and C-Classifier to predict similarity and location.

**C-RPN:** Alg 1 shows the detailed process of C-RPN. The feature fusion module consists of step 1-3. The learnable metric in C-RPN is implemented with two convolutional layers (step 4 in Alg 1), followed by two convolutional layers for predicting similarity scores and locations $P(p_i, bbox_i|I_q)$

**C-Classifier:** The entire workflow of C-Classifier is illustrated in Alg 2. Step 1-2 represents the feature fusion module. Here, the learnable metric (step 3-4 in Alg 2) consists of a convolutional layer for dimensionality reduction and two fully connected layers. After this, two fully connected layers computes similarity scores and locations $P(p_i, bbox_i|I_q)$.

**Algorithm 1 Framework of C-RPN.**

*Input:* Target features $\phi(I_t)$, Query features $\phi(I_q)$

*Output:* Similar regions $P(p_i, bbox_i|I_q)$

1. Generating global average- and max-pooling query features $GAP(\phi(I_q))$ and $GMP(\phi(I_q))$
2. Concatenating $GAP(\phi(I_q))$, $GMP(\phi(I_q))$ and halving dimensions to salient query features $S(\phi(I_q))$
3. Concatenating $S(\phi(I_q))$ to each position of $\phi(I_t)$ to fusion features
4. Computing similarity score $p_i$ and location $bbox_i$
5. return $P(p_i, bbox_i|I_q)$

**Algorithm 2 Framework of C-Classifier.**

*Input:* Target features $\phi(I_t)$, Query features $\phi(I_q)$, C-RPN prediction $P(p_i, bbox_i|I_q)$

*Output:* Similar regions $P(p_i, bbox_i|I_q)$

1. Extracting local target features $\phi^i(I_t)$
2. Resizing $\phi^i(I_t)$ and concatenating them
3. Halving dimensions of concatenated features
4. Computing similarity score $p_i$ and location $bbox_i$
5. return $P(p_i, bbox_i|I_q)$

The loss function is the same as in Faster R-CNN [8]. We adopt cross-entropy loss for classification and smooth $L_1$ loss for bounding box regression and calculate losses on many selected anchors. Non-Maximum Suppression (NMS) is used to reduce redundant detection boxes.

### 4. Dataset

Datasets used in previous related works [22, 23] have quite few object classes (usually only one). Their sizes are small (fewer than one thousand images), and their scenarios are relatively simple. These characters make previous datasets not suitable for deep learning based methods. In this paper, we build two benchmarks based on Fashion-MNIST [50] and PASCAL VOC [51], which contain bigger data size, more classes and complex scenarios.

**Fashion-OSCD:** Fashion-OSCD is a relatively simple dataset built on Fashion-MNIST inspired by [52]. This dataset is constructed by randomly scaling and embedding samples of Fashion-MNIST dataset into a larger image, and adding some noises with pieces of other samples (Figure 4). 7 classes are randomly selected as seen classes for training/validation, the other 3 unseen classes for test. Overall, we generate 8000 training, 8000 validation and 6999 test images.
Figure 4. Examples for Fashion-OSCD dataset. Green boxes denote ground truth regions. Each image contains as much as 3 objects with multi scales and aspect ratios. Noise is added to each image.

We randomly sample 6000 pairs per classes from the training subset, resulting in 42000 training pairs. Also, a query image is selected for each validation/test image randomly, so we can get 8000 validation pairs for each seen class and 6999 test pairs for every unseen class.

Pascal-OSCD: Moreover, we propose another relatively complex dataset called Pascal-OSCD from PASCAL VOC 2007 and 2012. We divide all 20 classes with a ratio of 4:1 so that there are 16 seen classes for training/validation and 4 unseen classes for test. Query images are cropped from object bounding boxes and paired with target images which contain the same class objects randomly. Training image pairs are sampled from the train&val subsets of Pascal VOC 2007 and 2012 datasets, while the validation/test pairs are generated from the test subset of Pascal VOC 2007 dataset. Thus, we have 100000 training, 6870 validation pairs and 762 test pairs for the Pascal-OSCD dataset. We will remove image pairs where the query object is cropped from the target image at test-time. This dataset is not difficult for supervised detection methods but brings challenging situations for OSCD, such as viewpoint or intra-class variation, occlusion, and partially visible for query object.

5. Experiments

We evaluate ComprisonNet and baselines on two proposed datasets in the following experiments. Although the transfer-learning low-shot detection methods are different from OSCD in data organization and approaches of training and test, we also evaluate them for qualitative comparison in 5.6.

5.1. Baselines

Though OSCD is valuable, there are quite few works on this topic due to technique limitation before. Previous works on OSCD adopt handcrafted features in their pipelines. Since numerous experiments [5, 8, 24, 25, 26, 27] have proven that deep models have remarkable advantages over handcrafted features, methods based on handcrafted features are out of our consideration naturally. We want to choose more competitive methods based on deep learning as baselines. Visual tracking can be viewed as an OSCD problem in principle, so we select two state-of-the-art works on visual tracking as baselines, and make some modifications to adapt them for one-shot condition detection task.

SiamFC[40]: It outputs a similarity score map by cross correlation between query and target features. Regions with similarity higher than a threshold are as the desired objects. We train this model as in [40] and set threshold = (meanscore + maxscore) / 2 to extract similar regions after testing many different thresholds.

SiamRPN[39]: SiamRPN performs one-shot detection by restricting the anchor positions within a narrow neighborhood of the pervious target position. We extend the anchor positions to the entire image and adopt it for OSCD.

5.2. Experimental Settings

We resize the input image to keep the length of its shortest side as a fixed value L. For query and target images, L = 84, 280 on Fashion-OSCD dataset and 100, 600 on Pascal-OSCD dataset. Anchor scales are [48, 96, 128] on Fashion-OSCD dataset and [128, 256, 512] on Pascal-OSCD dataset. Anchor aspect ratios are [1:1, 1:2, 2:1] on both Fashion-OSCD and Pascal-OSCD datasets.

We adopt AlexNet [24] as a backbone for all methods, and remove all fully connected layers on the relatively simple Fashion-OSCD dataset. For Pascal-OSCD dataset, VGG16 [25] is used as a backbone for all methods and remove the last pool layer and all the fully-connected layers. The whole framework is optimized by the loss function with SGD. Learning rate is from $10^{-3}$ to $10^{-6}$, epoch=10 on Fashion-OSCD and 30 on Pascal-OSCD.
5.3. Performance

For Fashion-OSCD dataset, ComparisonNet achieves 73.3% mAP on seen classes and 73.6% mAP on unseen classes (Table 1), gaining a significant improvement against two baselines, indicating the effectiveness of comparison learning. We can also see that performance varies from class to class. All models can achieve a high AP for trouser and bag because they are visually distinguishable, while AP for shirt and t-shirt are relatively low due to their similarity in appearance. The results show the difficulty for OSCD even on this simple dataset. Some detection results are shown in Figure 5.

Table 2 shows the results on Pascal-OSCD dataset. The proposed method yields 52.7% mAP on seen classes and 52.1% mAP on unseen classes, gaining a significant improvement against SiamFC and SiamRPN as well. The performance on Pascal-OSCD dataset is lower than Fashion-OSCD dataset mainly because of more complex scenarios and intra-class variation.

5.4. Two-Stage Pipeline

We execute conditional probability estimation in a two-stage pipeline to alleviate the difficulties of OSCD. On Fashion-OSCD dataset (Table 1), C-RPN achieves better performance on unseen classes (61.7% mAP) than SiamFC and SiamRPN, and it also outperforms SiamRPN on seen classes. The C-Classifier has an extra 23.4% and 11.9% mAP improvement on seen and unseen classes. Additionally, for Pascal-OSCD dataset (Table 2), model with only C-RPN achieves 27.4% mAP on seen classes and 32.1% mAP on unseen classes, yielding better performance than SiamFC and SiamRPN both on seen and unseen classes.

5.5. Visualization

In Figure 6 (a), we visualize some detection results on Pascal-OSCD dataset for qualitative analysis. The query image is at the top-right, and the target image is at the bottom. Partial occlusion and image blurring in query images makes model difficult to obtain the information of query images effectively. ComparisonNet detects objects of unseen classes successfully although there are considerable viewpoint and intra-class variation (tile 1,2,3,5,8 in Figure 6 (a)), furthermore, overcome distractions from seen classes such as car, person, bottle and tv (tile 3,4,12,15 in Figure 6 (a)), indicating the effectiveness of the learnable metric.

We visualize some hard examples and show query images, results of SiamFC, SiamRPN and ComparisonNet from left to right in each group of Figure 6 (b). All of
these models may not always acquire good enough query images due to random selecting, increasing the difficulties of OSCD. In group 3, 5 of Figure 6 (b), colors of query objects (cow and cat) are more similar to the distractors’ (person and tv) in target images. The query image in group 4 can only provide little information. And there are some distractors (person, car, tv) in group 1, 2, 3, 5. The proposed ComparisonNet performs better than other methods in these exceptionally difficult situations. No location regression is the fatal drawback of SiamFC, resulting in bad bounding boxes. SiamRPN is unable to work well when face the distraction of seen class objects.

5.6. Comparison with Transfer-learning Methods

(1) LSTD: We train and finetune LSTD as in [15].

(2) FRCNN: Faster R-CNN is trained and finetuned without extra regularizations.

(3) FRCNN\textsubscript{std}: Faster R-CNN is trained and finetuned with the two regularizations proposed in LSTD, i.e., BD (background-depression) and TK (transfer-knowledge).

For qualitative comparison, we train LSTD, FRCNN with seen class samples from VOC07+12 (the union set of VOC 2007 train & val and VOC 2012 train & val) in one-shot setting and evaluate it with seen class samples of VOC 2007 test dataset. Then, the trained model is finetuned with unseen class samples from VOC07+12 and evaluated with unseen class samples from VOC 2007 test dataset. For FRCNN\textsubscript{std}, the trained FRCNN is also finetuned with regularizations of LSTD and evaluated in same settings.

For seen classes, the transfer-learning methods are standard supervised learning detectors with sufficient training samples and achieve better performance than OSCD methods, but suffer from degrading performance on unseen classes, forming a striking contrast with the proposed OSCD framework.

5.7. Feature Fusion Methods

In order to get simple and effective salient query features for C-RPN, we evaluate three global pooling operations: max-, avg- and max&avg-pooling. For max&avg-pooling, we get salient query features $S(\phi(I_q))$ by concatenating the global max-pooling query features to global average-pooling query features, then halve its dimension by a convolutional layer. The evaluation results of 3 different feature fusion methods are shown in Table 3. Global max&avg-pooling is the best choice.

6. Conclusion

This paper proposes a novel OSCD framework based on the conditional probability theory, which aims to detect query objects from unseen classes without further training on these newly interested classes. Guided by the Bayes equation, we adopt a Siamese network to extract features for query and target images as the marginal probability. As the existence of strong non-linearity, a carefully designed C-RPN and C-Classifier have been implemented to form
a learnable metric approaching the solution of conditional probability under a coarse-to-fine pipeline, which refers the intended objects belong to the newly interested classes. In both detectors, our model learns to classify and locate query class objects by comparison learning. The proposed method may be extended to few-shot object detection easily by merging the features of increased samples across the query branch following the similar work in [14]. Experiments on two benchmarks based on the proposed Fashion-MNIST and PASCAL VOC dataset verify that our method achieves state-of-the-art performance for one-shot object conditional detection.

References

[1] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” International Conference on Learning Representations (ICLR), 2014.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Computer Vision and Pattern Recognition (CVPR), 2014.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in Computer Vision and Pattern Recognition (CVPR), pp. 779–788, 2016.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, “SSD: Single Shot MultiBox Detector,” in European Conference on Computer Vision (ECCV), pp. 21–37, 2016.

[5] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. PP, no. 99, pp. 2999–3007, 2017.

[6] R. B. Girshick, “Fast R-CNN,” International Conference on Computer Vision (ICCV), pp. 1440–1448, 2015.

[7] P. Purkait, C. Zhao, and C. Zach, “SPP-Net: Deep absolute pose regression with synthetic views,” Computer Vision and Pattern Recognition (CVPR), 2017.

[8] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” in International Conference on Neural Information Processing Systems, pp. 91–99, 2015.

[9] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” ICML’17 Proceedings of the 34th International Conference on Machine Learning - Volume 70, 2017.

[10] S. Ravi and H. Larochelle, “Optimization as a model for few-shot learning,” in In International Conference on Learning Representations (ICLR), 2017.

[11] Y. X. Wang and M. Hebert, “Learning to Learn: Model regression networks for easy small sample learning,” in European Conference on Computer Vision (ECCV), pp. 616–634, 2016.

[12] T. Munkhdalai and H. Yu, “Meta networks,” ICML’17 Proceedings of the 34th International Conference on Machine Learning - Volume 70, 2017.

[13] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, “Matching networks for one shot learning,” Advances in Neural Information Processing Systems 29 (NeurIPS), 2016.

[14] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, “Learning to compare: Relation network for few-shot learning,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[15] H. Chen, Y. Wang, G. Wang, and Y. Qiao, “LSTD: A low-shot transfer detector for object detection,” Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[16] M. Ren, S. Ravi, E. Triantafilou, J. Snell, K. Swersky, J. Tenenbaum, H. Larochelle, and R. S. Zemel, “Meta-learning for semi-supervised few-shot classification,” International Conference on Learning Representations (ICLR), 2018.

[17] A. A. Rusu, D. Yao, J. Sygnowski, O. Vinyals, and R. Hadsell, “Meta-learning with latent embedding optimization,” arXiv: Learning, 2018.

[18] A. Nichol, J. Achiam, and J. Schulman, “On-first-order meta-learning algorithms,” arXiv: Learning, 2018.

[19] T. Munkhdalai and A. Trischler, “Metalearning with hebbian fast weights,” arXiv: Neural and Evolutionary Computing, 2018.

[20] J. Snell, K. Swersky, and R. S. Zemel, “Prototypical networks for few-shot learning,” Advances in Neural Information Processing Systems 30 (NeurIPS), 2017.

[21] W. Yong, X. M. Wu, Q. Li, J. Gu, W. Xiang, Z. Lei, and V. O. K. Li, “Large margin few-shot learning,” arXiv: Learning, 2018.

[22] H. J. Seo and P. Milanfar, “Training-free, generic object detection using locally adaptive regression kernels,” IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 32, no. 9, p. 1688, 2010.

[23] S. K. Biswas and P. Milanfar, “One shot detection with laplacian object and fast matrix cosine similarity,” IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 38, no. 3, pp. 546–562, 2016.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in International Conference on Neural Information Processing Systems, pp. 1097–1105, 2012.

[25] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, “Very deep convolutional networks for text classification,” arXiv: Computation and Language, pp. 1107–1116, 2016.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2015.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” Computer Vision and Pattern Recognition (CVPR), pp. 1–9, 2014.
[28] X. Dong, L. Zheng, F. Ma, Y. Yang, and D. Meng, “Few-example object detection with model communication,” IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. PP, no. 99, pp. 1–1, 2017.

[29] S. Korman, D. Reichman, G. Tsur, and S. Avidan, “Fast-match: Fast affine template matching,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[30] I. Talmi, R. Mechrez, and L. Zelnik-Manor, “Template matching with deformable diversity similarity,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[31] Y. Yang, Z. Lu, and G. Sundaramoorthi, “Coarse-to-fine region selection and matching,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[32] G. Tolias, R. Sicre, and H. Jgou, “Particular object retrieval with integral max-pooling of cnn activations,” Computer Science, 2015.

[33] A. Gordo, J. Almazn, J. Revaud, and D. Larlus, “End-to-end learning of deep visual representations for image retrieval,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[34] K. Lin, H. F. Yang, J. H. Hsiao, and C. S. Chen, “Deep learning of binary hash codes for fast image retrieval,” in IEEE Conference on Computer Vision & Pattern Recognition Workshops, 2015.

[35] X. S. Wei, J. H. Luo, J. Wu, and Z. H. Zhou, “Selective convolutional descriptor aggregation for fine-grained image retrieval,” IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, vol. 26, no. 6, p. 2868, 2017.

[36] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” in Computer Vision & Pattern Recognition (CVPR), 2010.

[37] C. Engel, P. Baumgartner, M. Holzmann, and J. F. Nutzel, “Person re-identification by support vector ranking,” in British Machine Vision Conference (BMVC), 2010.

[38] H. Yao, S. Zhang, D. Zhang, Y. Zhang, J. Li, W. Yu, and T. Qi, “Large-scale person re-identification as retrieval,” in IEEE International Conference on Multimedia & Expo, 2017.

[39] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, “High performance visual tracking with siamese region proposal network,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8971–8980, 2018.

[40] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, “Fully-convolutional siamese networks for object tracking,” in European Conference on Computer Vision (ECCV), pp. 850–865, 2016.

[41] W. Chong, W. Ren, K. Huang, and T. Tan, Weakly Supervised Object Localization with Latent Category Learning, European Conference on Computer Vision (ECCV), 2014.

[42] E. W. Teh, M. Rochan, and Y. Wang, “Attention networks for weakly supervised object localization..” British Machine Vision Conference (BMVC), 2016.