Detecting Objects in Less Response Time for Processing Multimedia Events in Smart Cities

Asra Aslam
Insight Centre for Data Analytics
NUI Galway, Ireland
asra.aslam@insight-centre.org

Abstract

Due to increase in multimedia traffic in smart cities, we are facing the problem of processing unseen classes in real-time. Existing neural-network based object detectors may support this growing demand of multimedia data but have the limitation of availability of trained classifiers for unseen concepts. This results in a long waiting time for users who want to detect unseen classes. In this paper, we proposed three approaches where we can utilize existing object detection models and can train unseen classes within short training time. Our approaches are based on similarity of unseen classes with seen classes, and availability (presence or absence) of bounding boxes. Our results indicate that the proposed framework can achieve accuracy between 95.14% to 98.53% within response time of ∼0.01 min to ∼30 min for seen and partially unseen classes. Moreover we achieve state of the art results (68.78 mAP within 10 min) for unseen classes that have only image-level labels for training and no bounding boxes. Our qualitative results indicate that our approaches can work well for any unseen class (not only for conventional object detection datasets).

1. Introduction

Event-based multimedia approaches exhibit high performance in the current scenario but are designed for specific domains (like traffic management, security, supervision activities, terrorist attacks, natural hazards [15, 21, 30, 39, 53]) and hence can handle only familiar classes (have bounded/limited vocabulary). The escalating growth of multimedia data with large numbers of user subscriptions poses multiple challenges for the processing of image-based events in smart cities [2]. On the other hand, a large number of unseen concepts are emerging and changing over time in various domains in smart cities. Furthermore, the essential requirement of multimedia applications is a real-time performance [1,36], which needs to be fulfilled for its usability. This highlights the need for minimization of *response time* while maintaining *accuracy* from the user's perspective for new/unseen classes. However, the existing object detection approaches including few shot models [26, 27, 35, 41, 51] takes long training time (days or large number of resources) to train on new classes.

Consider the scenario of object detection for analyzing image based events in smart cities (shown in Fig. 1). Suppose a user subscribes for the detection of "Bus" on "Bus Stand". This type of query can be answered "Public Transport Management" using a camera observing the bus stand and producing images consisting of bus-status related information. Similarly, if a user subscribes for the detection of the empty parking spot (i.e. absence of car at parking spot), we require another application for processing "Car Parking Management" events. Moreover, if a user subscribes to the concepts like "taxi" or "pedestrian", then existing public transport and car parking management systems will not be able to respond to any new class even if they already consist of similar classes like "car" and "person". Thus, we need a generalizable image event processing system that can provide adaptation from seen to unseen concepts (like *car* to *taxi*) and able to answer any completely unseen concept (like a *cat*, *dog*, *key*, *bicycle*, etc.) of any domain. Existing deep neural network based model can be very useful if they can be trained in short time for unseen concepts.

In this paper we proposed three approaches that can be used in different scenarios of training models for unseen classes. We summarize the main contributions as follows:

- Hyperparameter tuning based approach for completely new classes which requires training from scratch.
- Domain adaptation based approach for partially unseen classes where similar seen classifiers are available.
- Classifier to Detector conversion based approach for partially unseen classes which have some similarity with seen classes but does not have bounding boxes.

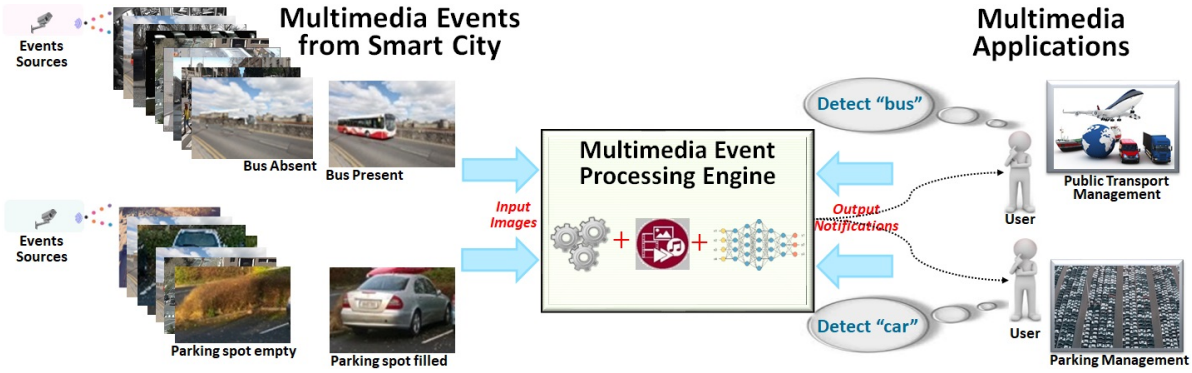


Figure 1. Generalizable Multimedia Event Processing for the Detection of Unseen Classes in Smart Cities.

2. Related Work

Adaptive classifiers and Self-Tuning Existing adaptive classifier based machine learning techniques [11, 13, 49, 58] in this category are designed with the aim of evolution of classifiers with drift in concept of multimedia streams. The identification of *concept drift* in these dynamic approaches, is mainly focused on processing of text data streams and cannot accommodate multimedia data streams.

Auto-WEKA [46] is one of the most popular work towards analyzing machine learning algorithms automatically and setting appropriate hyperparameters in-order to enhance performance. Similarly hyperopt-sklearn is another available software mainly includes random search and TPE for the automatic selection [5]. Inspite of the fact that these tools are automatic, most of them focuses only on accuracy and generalization ability of classifiers, or on the computation cost consisting of testing time [19, 48], while excluding the training time of neural-network based models

Transfer learning based methods Many approaches [4, 40] with supervised/unsupervised transfer learning have been proposed for domain adaptation and mainly focused on the generalization ability for increasing accuracy not the overall response time. An event recognition in still images by transferring objects and scene representations has been proposed in [50], where the correlations of the concepts of object, scene, and events have been investigated. The authors proposed techniques to exploit the knowledge from other networks, and also evaluated the model on multiple event domains. Another domain adaptation approach based on the Faster R-CNN object detection model has been proposed [9] in order to reduce the domain discrepancy and enhance the effectiveness for cross-domain object detection. Such approaches are robust; however, in most cases, domain shift represents different changes in view-points, weather conditions, backgrounds, image quality, sketches, etc.

Weakly supervised learning Recently, weakly supervised learning [25, 31, 47, 57] is emerging as a possible solution for large-scale unseen concepts. Such approaches are designed for limited classes and cannot incorporate new classes that have no pre-trained models. Weakly supervised learning [56] also formulated as a Multiple Instance Learning (MIL) problem. Most of the existing approaches [10, 16, 42, 43] are evaluated on classes of Pascal VOC [12], disregard the training time, and/or use Fast RCNN [14] as base network. Similarly, large scale domain adaptation based methods [17, 44] are also introduced particularly for the detection of objects and it is desirable to bring their abilities to core of multimedia event processing.

3. Approaches for Unseen Class Object Detection

The problem of multimedia event processing is divided into different scenarios shown in Fig. 2. Suppose a user subscribes for a concept, if we find a classifier that can detect subscribed concept, we call the concept "Seen" and recognize it with "Scenario 0". In this case, we process the subscriptions directly using the existing classifier, without training any other model for the *seen* concept. However, if we don't find any classifier to process the unseen concept, then we attempt to find "Any similar seen concept available?". We use the names of seen classes and compute their individual similarities with the subscribed unseen class. In the worse case, if the concept is completely "unseen", we introduce "Scenario 1" for the handling of subscriptions that are not related to any domain and resulted in low similarity scores. On the occurrence of complete "unseen" concept, we train the classifiers from scratch and optimize the training by hyperparameter tuning to reduce the overall response time of the multimedia event processing model. The most likely scenario is to receive the concept which is "unseen" and have similarity with one or more "seen" concepts. Thus, we can train classifiers for such unseen domains by

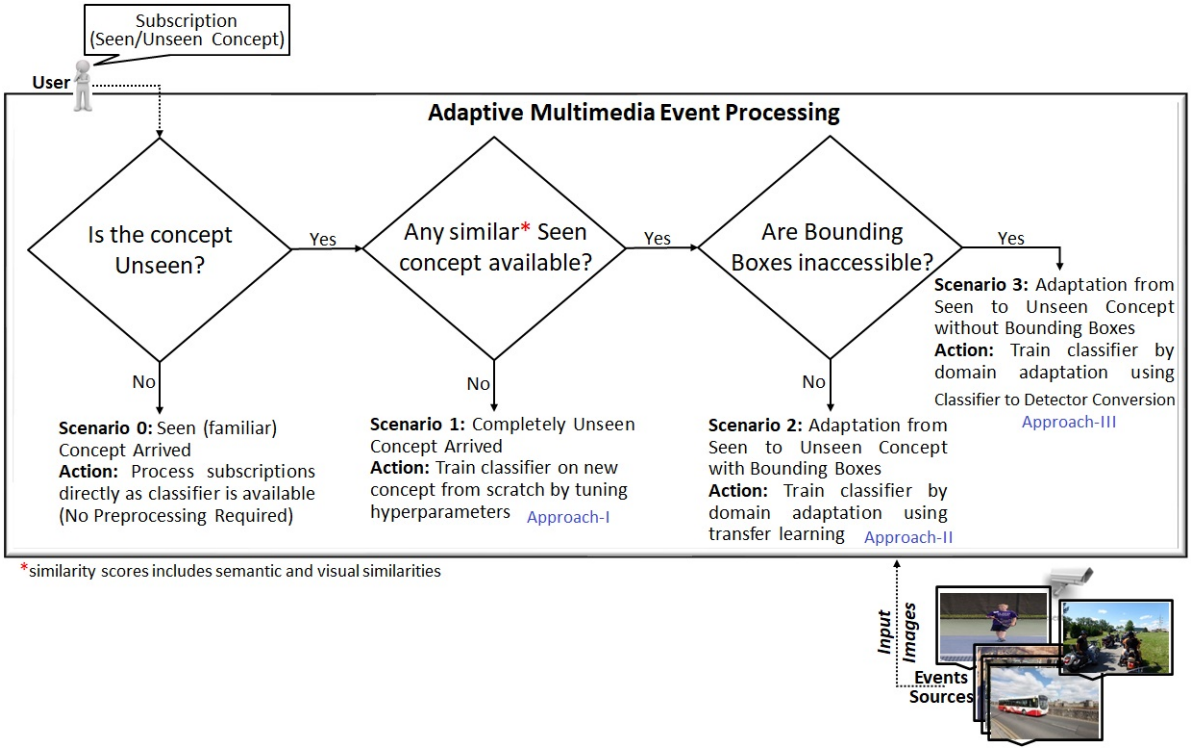


Figure 2. Scenarios for Detecting Unseen Objects in Multimedia Event Processing of Smart Cities.

knowledge transfer from seen domains (Scenario 2). Our final concern is the availability of bounding box annotations to train classifiers for subscribed concepts (Scenario 3). In this case, we use the classifier to detector conversion methods for the training of the detector for unseen classes.

3.1. Approach-I: Using Hyperparameters Tuning

Consider the scenario when the concept is completely "unseen", i.e., there is no similar seen concept-based classifier available in the multimedia event processing model for the knowledge transfer. This specific problem associated with Scenario-1 is described in Fig. 3. For instance, a user subscribes for *mirror* detection and existing multimedia event processing model can detect only *bicycle*, *bus*, *cat*, *traffic*, *person* etc. In that case, we need to address the problem of classifier training for unseen concepts from scratch in low response time.

In this Approach-I, we incorporated object detection model with self hyperparameter tuning to train for unseen concepts in the required response time. As the choice of hyperparameter values greatly affects the performance of resulting classifiers, we leverage self-tuning of optimization hyperparameters, including the configuration of learning-rate, batch-size, and the number of epochs for minimizing the training time while maintaining the promising accuracy. In this case, our model evaluates the training time and computes hyperparameters automatically based on prerequisite response time. Here, we train models from scratch as we assume no seen class-based classifiers are available, which is not true in most cases. Thus, we incorporate transfer learning to further improve performance in the next scenario.

3.2. Approach-II: Using Domain Adaptation

The problem associated with Scenario-2 is described in Fig. 4, where we need to reduce the response time for cases where the intended classifier is not available but similar classifiers are available. Suppose a user subscribes to the detection of class "car", unseen to the multimedia event processing model. If a model already consists of related classifiers (like *bus* in the present case), we need to train classifiers for such partial unseen concepts in a reduced response time.

In Approach-II, on arrival of new concept, the proposed model first identifies if any similar classifier is available, or if there is any possibility for domain adaptation. Second, it performs the training of classifiers on need for the intended new subscription. More specifically for domain adaptation, we are using fine-tuning and freezing of classifier layer based methods. In this approach, we use the previously trained classifier to instantiate the network of

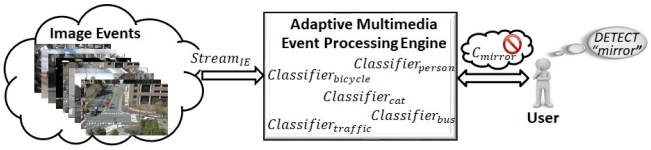


Figure 3. Scenario-1: Completely Unseen Concept Arrived (Approach-I: Hyper-Parameters Tuning for short Training Time)

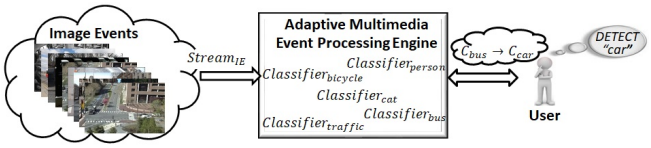


Figure 4. Scenario-2: Partially Unseen Concept Arrived (Approach-II: Domain Adaptation for short Training Time)

another classifier required for a similar subscription concept. We freeze the backbone (convolutional and pooling layers) of the neural network and train only top dense fully connected layers. As frozen backbone is not updated during back-propagation and only fine-tuned layers are updated and retrained during the training of classifier, this results in less training time with reasonable accuracy. Presently we are using the *path vector* operator as a WordNet relatedness measure [32] for the computation of similarity among subscriptions, which could be replaced in future with more accurate measures using image-feature based domain-specific ontologies depending on the utility of applications.

3.3. Approach-III: Using Classifier to Detector Conversion

The last scenario (in Fig. 2) is also associated with the partial unseen concept, but it removes the limitation of the previous Scenario-2, where we need annotated bounding boxes to train models. Presently, most of the object detection datasets have limited vocabulary; thus, we cannot provide bounding boxes for a large number of unseen concepts.

Suppose a user subscribes for an unseen class "dog" and we have only image-level labels available for training. However, we have another seen class "cat" which is visually and/or semantically similar to *dog* class and we have image-level as well as object level labels for it. Then we can train cat classifier and cat detector using them. Moreover, we can train dog classifier using its available image-level labels. Then using the last layer weights differences between cat classifier and detector we can create dog detector, by adding that weight difference to the last layer of dog detector. The demonstration of example is shown in Fig. 5.

We propose an "unseen class model", which allows a

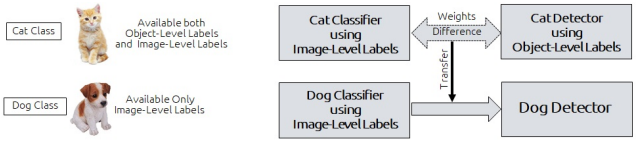


Figure 5. Classifier to detector conversion for Unseen classes (like dog) that has only image-level labels and no bounding boxes.

Table 1. Hyperparameter values for Adaptive Training

Hyperparameter for different Training Time	Model	Batch Size	Learning Rate	#Epochs
Defaults Hyperparameters for long training time	YOLO	64	0.001	300
	SSD	32	0.001	120
	RetinaNet	1	1e-5	50
Our Derived Hyper-parameters for short response time (<15 min)	YOLO	64	0.005315	2
	SSD	8	0.002612	2
	RetinaNet	1	0.000195	5
Our Derived Hyper-parameters for average response time (<60 min)	YOLO	64	0.007935	9
	SSD	4	0.003600	12
	RetinaNet	2	0.000224	9

user to construct detectors for unseen classes without the need for detection data (no bounding boxes) within the short training time. Our model is based on making use of existing object detection datasets of bounded vocabulary (consists of *seen* concepts) to construct detectors for *unseen* concepts (i.e., unbounded vocabulary) by using the differences between a weak detector (trained on image classification dataset) and a strong detector (trained on object detection datasets).

In Approach-III, we train two separate detectors, one on existing object level labels (of MCOCO, and OID dataset) and another on image-level labels (using ImageNet dataset), respectively. Then Approach-III follows the below steps:

1. Download images using only image-level labels on request of any *unseen* concept (like dog).
2. The *object-level detector* is then fine-tuned on collected images of unseen concepts by labeling the most semantically similar class (like cat) with the unseen class name (like dog).
3. At this stage, we compute the visual similarity of the constructed unseen class detector (trained on classification data) with seen classes of *image-level detector*, combine it with semantic similarities, and select top-k classes ranked on comprehensive similarities. Visual similarity is presently the difference between the weights of the last layers of seen and unseen classes.
4. We transfer the knowledge of classifier-detector differences of top classes to the constructed unseen class detector and adapt it into the stronger detector without further training.

Table 2. Comparison of performance of Approach-2 for short training time using default and derived hyperparameters

Hyperparameters*	**mAP on low Response Time** ($15\tilde{m}in$)			**mAP on average Response Time** ($60\tilde{m}in$)		
	YOLO	SSD	RetinaNet	YOLO	SSD	RetinaNet
Using Default Hyperparameters	0.00	0.06	0.13	0.09	0.00	0.20
Using Ours Hyperparameters for Approach-I	0.01	0.03	**0.20**	0.10	0.00	**0.32**

* Please see Table–1 for Hyperparameter Values.

5. Finally, we perform the detection using our trained network communicate results.

In this approach all networks use YOLOv3 with MobileNetv3 backbone as our aim is to reduce response time and these are fast models for detection and classification.

4. Experiments

4.1. Implementation Details

Here, we utilized Pascal VOC, Microsoft COCO, and Open Images dataset (OID) for the construction of detectors. Specifically for Approach I, number of training images for the subscriptions cat, dog, laptop, car, bus, bicycle, and football classes are 1804, 2204, 5528, 2820, 847, 1108, and 4339. If bounding box annotations of image consist of any of the classes (cat, dog, laptop and so on), then added it to testing events set. The number of testing events for the same classes are 384, 538, 355, 1588, 256, 396, and 413 respectively. In Approach II, we added classes cricket ball, laptop bus, and mango classes consist of 95, 5528, and 126 training images; and 15, 355, and 23 testing images respectively.

In Approach III, object-level detector consist of 80 classes of Microsoft COCO [28] and 20 classes of OID [23]; while image-level detector consist of same 100 seen classes of ISLVRC [37] dataset. We chose unseen classes in such a way that the same classes should be present in OID (consist of 600 classes). So that testing dataset of OID can be served as groundtruth. For the training of 100 unseen classes, constant learning rate of 10^{-4} is used and evaluated on 0.5 IoU.

4.2. Comparison of performance before and after adaptation for short response time

We performed our investigation on different object detection models (YOLOv3 [34], SSD-300 [29], and RetinaNet [27]). Table 1 represents default hyperparameter values and our derived hyperparameter values for short training time. We derived the values for low and average training time by performing experiments on large number of trials and using TPE search method [6], which needed to be minimized based on mean average precision (mAP).

It can be observed in Table 2 that in most of the case our approach based hyperparameters gives us better mAP than default hyperparameters of object detection models. Here we took 15 min and 60 min as an example to represent less and average training time respectively. Using Approach I, we found that RetinaNet performs best on low and average training time.

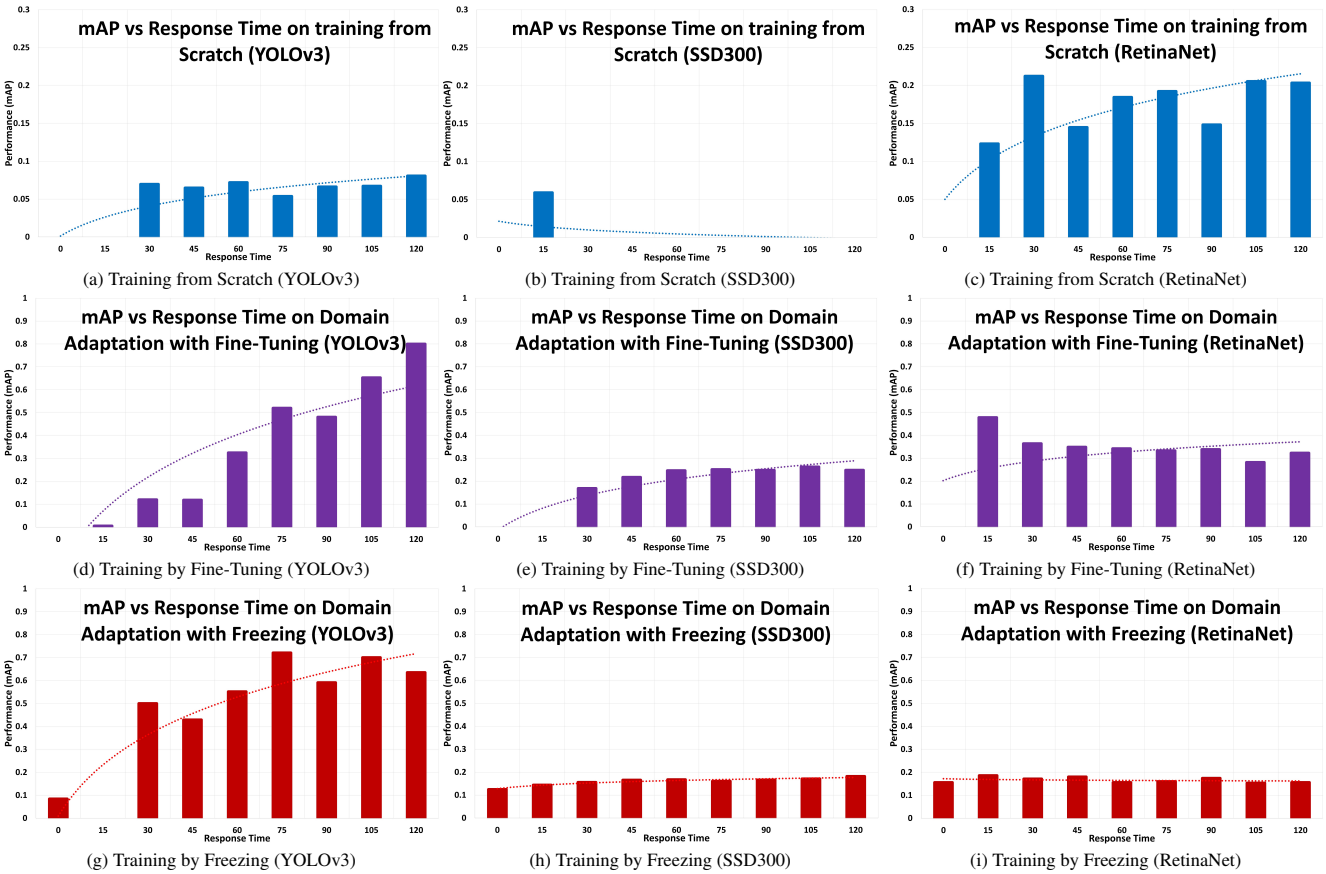
4.3. Comparison of performance for domain adaptation for short response-time

We evaluate transfer learning techniques on same models, to analyze which classifiers can perform well on applying what type of training (scratch, fine-tuning, and freezing layers). The results of performance with *response time* trade-off are shown in Fig. 6. Firstly, it represents the trade-off on arrival of a completely new subscription, when there is no possibility of domain adaptation, for the training time of 120 min. In this case, all models are trained from scratch without the use of any pre-trained model. We can observe the performance of RetinaNet (Fig. 6c) is higher than other object detection models and the SSD model (Fig. 6b) is very difficult to converge with training from scratch, thus resulting in the worse performance, whereas the performance of YOLOv3 (Fig. 6a) is also low. However, by choosing training time ($\sim$30 min) using RetinaNet, we can reach accuracy $\sim 77.10\%$ with precision ~ 0.21.

The performance of RetinaNet and SSD are better than YOLOv3 in initial 30 min of training for both cases of fine-tuning (Fig. 6d, 6e, and 6f) and freezing (Fig. 6g, 6h, and 6i) layers. However there is a sudden rise in performance of YOLOv3 in the first few minutes signifies its higher slope in terms of short time training as compared to other object detection models. Trend lines are also shown with comparison, just to give the clear demonstration of initial precision that a particular object detection model can achieve at zero response time, as well as the highest precision of a model in a given response time. We can easily observe that all object detection models with fine-tuning technique are performing better as compared to the adaptation technique of freezing layers for long training time ($>$ 120 min). However for short training time ($<$ 30 min) YOLOv3 with freezing technique and RetinaNet with fine-tuning are performing best.

It is worth to mention that the testing time on our resources for YOLOv3, SSD300, and RetinaNet are 0.009 sec, 0.05 sec, and 0.08 sec for one image.

Table 3 shows four examples of classes on domain transfers with different similarity scores. It also presents *Transfer Loss*, *Accuracy*, and *Distribution Discrepancy* metrics computed for analyzing domain transfers (*laptop* to *mango*,



(a) Training from Scratch (YOLOv3) (b) Training from Scratch (SSD300) (c) Training from Scratch (RetinaNet)

(d) Training by Fine-Tuning (YOLOv3) (e) Training by Fine-Tuning (SSD300) (f) Training by Fine-Tuning (RetinaNet)

(g) Training by Freezing (YOLOv3) (h) Training by Freezing (SSD300) (i) Training by Freezing (RetinaNet)

Figure 6. Performance vs Response Time with and without Domain Adaptation

Table 3. Analysis of Domain Adaptation using Transfer Loss, Accuracy, and A-Distance

Classes with Domain Adaptation	Semantic Similarity	Transfer Loss			Accuracy			A-Distance		
		YOLO	SSD	RetinaNet	YOLO	SSD	RetinaNet	YOLO	SSD	RetinaNet
Mango Detector from Laptop Detector	0.08	0.00%	14.21%	13.73%	93.26%	66.07%	56.31%	1.08	1.12	1.70
Dog Detector from Cat Detector	0.20	0.24%	-0.97%	-7.25%	90.43%	77.83%	74.27%	0.76	1.59	1.59
Cricket_Ball Detector from Foot_Ball Detector	0.33	0.25%	13.87%	23.89%	95.70%	65.52%	51.02%	1.04	1.12	1.65
Bus Detector from Car Detector	0.50	2.96%	-3.51%	1.31%	95.33%	71.66%	63.73%	1.01	1.35	1.68

Football to *cricket ball*, *car* to *bus*, and *cat* to *dog*). Transfer loss indicates how well the transfer works on multiple domains, and its lower values are recommended. In this case the best transfer is achieved by RetinaNet on the transfer of *cat* to *dog* class. However YOLOv3 achieve the best accuracy on all domain transfers. In-order to realize the variation of approximate distance (i.e. Distribution Discrepancy) among different domains, we have trained few binary classifiers that can classify source-target pair of classes. Here also YOLOv3 neural network closes the cross-domain gap more effectively as distribution discrepancy (lower is better) is lowest with it as compared to other object detection models. The qualitative detection performance after applying Approach-II on unseen class "bus" on an event of bus arrival and leaving of bus on bus stand is shown in Fig. 7.

Table 4 provides a comparison of average accuracy and response time of proposed with existing models by considering their best performance. It can be observed that existing domain specific models are designed only for the detection of specific objects and answer such seen (known) subscriptions in low response time, however they fail to process any unseen (unknown) subscription of different domain.

(a) Frame-1 (b) Frame-2 (c) Frame-3 (d) Frame-4 (e) Frame-5 (f) Frame-6 (g) Frame-7 (h) Frame-8

Figure 7. Example frames of detection of presence/absence of unseen class "bus" at bus stand using Approach 2.

Table 4. Comparison of Proposed with Existing Model(s)

Approach	Example of Seen/Unseen Classes		Performance	
			Response Time	Accuracy
Existing Domain Specific Models	Vehicle Detection [45] (Seen)		0.001 min	97.30%
	Firearm Detection [24] (Seen)		0.0001 min	94.00%
	Stolen Objects [38] (Seen)		0.0007 min	93.58%
	Car Parking Vacancy [20] (Seen)		0.17 min	97.90%
	Traffic Light, Key, Pedestrian, Ball etc. (Unseen)		∞	0.00%
Proposed Approaches	Car, Football, Cat, Laptop etc. (Seen)		0.01 min	98.53%
	Bus, Dog, Mango, Cricket ball, Mirror etc. (Unseen)	Approach-I	15 min (low)	79.00%
		Approach-I	60 min (average)	84.28%
		Approach-II	30 min	95.14%
		Approach-III	10 min (no BBox)	42.86%

The proposed model can achieve accuracy of 95.14% even when all concepts are unseen by taking an average response time of ∼30min. Also, the accuracy with Approach-I on low and average response time are 79.00% and 84.28% respectively. Moreover, with Approach-III we get 42.86% accuracy but without bounding boxes, and with training time of only 10 min (discussed in next section).

4.4. Analysis of Performance for domain adaptation without bounding boxes for short response time

To retrieve the effective range of response-time in our model, we train each category until the point testing accuracy starts to decrease (to avoid overfitting). We show a few examples of unseen concepts in Figure 8. Please note here we compute the total number of epochs for varying the training time. We first train our model on weak level labels (i.e., without bounding boxes) and then test on strong labels (i.e., with bounding boxes). We observe that the maximum mAP of each class could be achieved within 10 min of training. After that, the mAP decreases and remains constant. We recommend 10 min of training to attain maximum mAP 43.07. It is worth noting that mAP at 0 min training time on unseen classes is not zero due to the use of detectors trained on object-level labels for initialization.

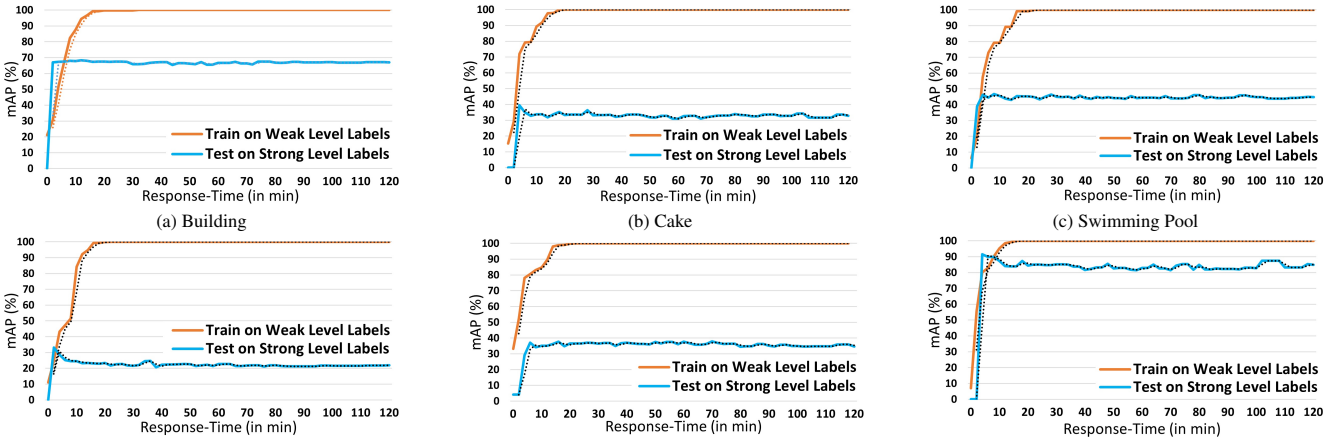
We present an analysis in Figure 10 of few *unseen* categories along with their respective degree of similarity with *seen* categories (top-10 nearest neighbor). Here, we compute comprehensive similarity scores using visual and semantic similarity. It can be observed that unseen class detectors performs well for most of the classes due to knowledge transfer from seen class detectors.

We show few examples of our model detections for qualitative analysis in Fig. 9. Here unseen classes of Fig. 9 (a) – (d) consist of OID dataset images where groundtruths are shown in green color while our detections are shown in red color. Fig. 9 (e) – (h) consists of additional unseen classes which are present neither in any object detection nor in image classification datasets to date. Due to this reason we have only our detections (in *red* color) for these images.

Presently, existing few-shot object detection models are showing great promise by providing competitive performance with only few shots of annotated. Thus, we compared our model performance with recent zero-shot, one-shot, and few-shot detection approaches in Table 5. Due to lack of space, in case of few shot object detection we show here the performance of 10 shots of existing approaches [8, 22, 26, 41, 51, 52, 54, 55] which is best among all. It is important to note that we use only image-level labels; thus, our approach does not need any shot.

5. Conclusion

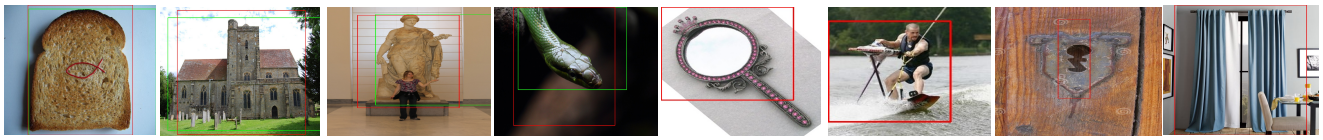
The problem of processing multimedia events that can include a large number of seen/unseen concepts belonging to the same or multiple domains are analyzed in this paper. We proposed approaches for completely unseen concepts, partially unseen concepts, and unseen concepts without bounding boxes. Our approaches utilized hyperparameter tuning, domain adaptation, and classifier to detector conversion method to training unseen classes detectors in short time. The proposed approach can achieve the accuracy ranges from 95.14% to 98.53% within $\sim$ 0.01 min to $\sim$ 30 min. Moreover, we achieve mAP 68.78 in training time of 10 min on unseen classes of without bounding boxes. Qual-



(a) Building (b) Cake (c) Swimming Pool

(d) Goat (e) Candy (f) Pasta

Figure 8. Examples of mAP with Response-Time, For each "Unseen" category, we use the top-10 weighted average nearest neighbor "Seen" categories for adaptation. This shows that maximum mAP could be achieved within 10 min.



(a) Bread (b) Building (c) Sculpture (d) Snake (e) Mirror (f) Ironing Board (g) Keyhole (h) Curtain

Figure 9. Examples of detections of our model on "Unseen" categories shown in red color and groundtruth (taken from OID) in green. Last four unseen classes images are downloaded online, so no groundtruth available to date.

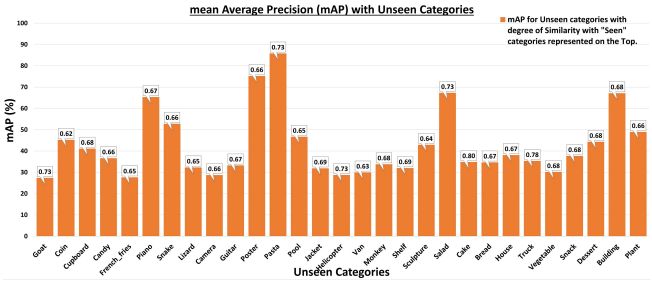


Figure 10. mAP of our model on 100 "Unseen" Categories within 10 min of training. We use the top-10 nearest neighbor "Seen"categories for each unseen category, with average similarity scores shown on top.

Methods	mAP
Zero-Shot [3, 33]	15.32
One-Shot [7, 18]	72.2
Few-Shot [8, 22, 26, 41, 51, 52, 54, 55]	57.37
Ours*	68.78

*Like existing methods, mAP is computed for random 5-class splits.

Table 5. Comparison with N-Shot detection models.

itative results demonstrates that our approaches are suitable for any unseen class. In future work, the model can be extended for unsupervised learning to reduce the need of labeled data for the processing of new subscriptions. The use of image features based domain-specific ontologies for the computation of similarity among domains could lead to the enhancement in performance.

Acknowledgments This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 P2, co-funded by the European Regional Development Fund.

References

[1] Asra Aslam and Edward Curry. Towards a generalized approach for deep neural network based event processing for the internet of multimedia things. *IEEE Access*, 6:25573–25587, 2018. 1

[2] Asra Aslam and Edward Curry. A survey on object detection for the internet of multimedia things (iomt) using deep learning and event-based middleware: Approaches, challenges, and future directions. *Image and Vision Computing*, 106:104095, 2021. 1

[3] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 8

[4] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36, 2012. 2

[5] James Bergstra, Dan Yamins, and David D Cox. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in Science Conference*, pages 13–20. Citeseer, 2013. 2

[6] J. Bergstra, D. Yamins, and D. D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages I–115–I–123. JMLR.org, 2013. 5

[7] Ding-Jie Chen, He-Yen Hsieh, and Tyng-Luh Liu. Adaptive image transformer for one-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12247–12256, 2021. 8

[8] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. Lstd: A low-shot transfer detector for object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 7, 8

[9] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018. 2

[10] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2016. 2

[11] Božidara Cvetković, Boštjan Kaluža, Matjaž Gams, and Mitja Luštrek. Adapting activity recognition to a person with multi-classifier adaptive training. *Journal of Ambient Intelligence and Smart Environments*, 7(2):171–185, 2015. 2

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 2

[13] Dewan Md Farid, Li Zhang, Alamgir Hossain, Chowdhury Mofizur Rahman, Rebecca Strachan, Graham Sexton, and Keshav Dahal. An adaptive ensemble classifier for mining concept drifting data streams. *Expert Systems with Applications*, 40(15):5895–5906, 2013. 2

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[15] Holger Glasl, David Schreiber, Nikolaus Viertl, Stephan Veigl, and Gustavo Fernandez. Video based traffic congestion prediction on an embedded system. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, pages 950–955. IEEE, 2008. 1

[16] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold mil training for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2409–2416, 2014. 2

[17] Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*, pages 3536–3544, 2014. 2

[18] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. *Advances in Neural Information Processing Systems*, 32, 2019. 8

[19] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*, volume 4, 2017. 2

[20] Jermsak Jermsurawong, Mian Umair Ahsan, Abdulhamid Haidar, Haiwei Dong, and Nikolaos Mavridis. Car parking vacancy detection and its application in 24-hour statistical analysis. In *2012 10th International Conference on Frontiers of Information Technology*, pages 84–90. IEEE, 2012. 7

[21] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Disaster monitoring using unmanned aerial vehicles and deep learning. *arXiv preprint arXiv:1807.11805*, 2018. 1

[22] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019. 7, 8

[23] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2:3, 2017. 5

[24] Mikolaj E. Kundegorski, Samet Akcay, Michael Devereux, Andre Mouton, and Toby P. Breckon. On using feature descriptors as visual words for object detection within x-ray baggage security screening. In *7th International Conference on Imaging for Crime Detection and Prevention (ICDP 2016)*, pages 1–6, 2016. 7

[25] Yan Li, Junge Zhang, Kaiqi Huang, and Jianguo Zhang. Mixed supervised object detection with robust objectness transfer. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):639–653, 2018. 2

[26] Yiting Li, Haiyue Zhu, Yu Cheng, Wenxin Wang, Chek Sing Teo, Cheng Xiang, Prahlad Vadakkepat, and Tong Heng Lee. Few-shot object detection via classification refinement and distractor retreatment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15395–15403, 2021. 1, 7, 8

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 5

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In

European conference on computer vision, pages 740–755. Springer, 2014. 5

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 5

[30] Pirkko Mustamo. Object detection in sports: Tensorflow object detection api case study. *University of Oulu*, 2018. 1

[31] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 685–694, 2015. 2

[32] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pages 38–41. Association for Computational Linguistics, 2004. 4

[33] Shafin Rahman, Salman H Khan, and Fatih Porikli. Zero-shot object detection: joint recognition and localization of novel concepts. *International Journal of Computer Vision*, 128(12):2979–2999, 2020. 8

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 5

[35] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016. 1

[36] Aluizio Rocha Neto, Thiago P Silva, Thais Batista, Flávia C Delicato, Paulo F Pires, and Frederico Lopes. Leveraging edge intelligence for video analytics in smart city applications. *Information*, 12(1):14, 2021. 1

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 5

[38] Juan Carlos San Miguel and José M Martínez. Robust unattended and stolen object detection by fusing simple algorithms. In *2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*, pages 18–25. IEEE, 2008. 7

[39] Chiao-Fe Shu, Arun Hampapur, Max Lu, Lisa Brown, Jonathan Connell, Andrew Senior, and Yingli Tian. Ibm smart surveillance system (s3): a open and extensible framework for event based surveillance. In *Advanced Video and Signal Based Surveillance, 2005. AVSS 2005. IEEE Conference on*, pages 318–323. IEEE, 2005. 1

[40] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, volume 6, page 8, 2016. 2

[41] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7362, 2021. 1, 7, 8

[42] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2843–2851, 2017. 2

[43] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 352–368, 2018. 2

[44] Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, and Liming Chen. Large scale semi-supervised object detection using visual and semantic knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2016. 2

[45] Yong Tang, Congzhe Zhang, Renshu Gu, Peng Li, and Bin Yang. Vehicle detection and recognition for intelligent traffic surveillance system. *Multimedia tools and applications*, 76(4):5817–5832, 2017. 7

[46] Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855. ACM, 2013. 2

[47] Jasper Uijlings, Stefan Popov, and Vittorio Ferrari. Revisiting knowledge transfer for training object class detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1101–1110, 2018. 2

[48] Jonas Vlasselaer, Wannes Meert, and Marian Verhelst. Towards resource-efficient classifiers for always-on monitoring. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 305–321. Springer, 2018. 2

[49] Haixun Wang, Wei Fan, Philip S Yu, and Jiawei Han. Mining concept-drifting data streams using ensemble classifiers. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 226–235. AcM, 2003. 2

[50] Limin Wang, Zhe Wang, Yu Qiao, and Luc Van Gool. Transferring deep object and scene representations for event recognition in still images. *International Journal of Computer Vision*, 126(2-4):390–409, 2018. 2

[51] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *International Conference on Machine Learning (ICML)*, 2020. 1, 7, 8

[52] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9925–9934, 2019. 7, 8

[53] Xiu-Shen Wei, Bin-Bin Gao, and Jianxin Wu. Deep spatial pyramid ensemble for cultural event recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 38–44, 2015. 1

[54] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *European Conference on Computer Vision*, pages 456–472. Springer, 2020. 7, 8

[55] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9577–9586, 2019. 7, 8

[56] Zhaoyang Zeng, Bei Liu, Jianlong Fu, Hongyang Chao, and Lei Zhang. Wsod2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8292–8300, 2019. 2

[57] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1841–1850, 2017. 2

[58] Alexei Zhukov, N Tomin, V Kurbatsky, Denis Sidorov, Daniil Panasetsky, and Aoife Foley. Ensemble methods of classification for power systems security assessment. *Applied Computing and Informatics*, 2017. 2