Weakly supervised learning of deformable part models and convolutional neural networks for object detection

Yuxing Tang



THÈSE

pour obtenir le grade de

DOCTEUR DE L'ECOLE CENTRALE DE LYON

Spécialité: Informatique

N° d'ordre NNT : 2016LYSEC62

Weakly Supervised Learning of Deformable Part Models and Convolutional Neural Networks for Object Detection

---

dans le cadre de l'Ecole Doctorale InfoMaths

présentée et soutenue publiquement par

Yuxing Tang

14 décembre 2016

Directeur de thèse: Prof. Liming Chen

Co-encadrant de thèse: Dr. Emmanuel Dellandréa

JURY

Dr. Francesc Moreno-Noguer	IRI, UPC	Rapporteur
Dr. Guoying Zhao	University of Oulu	Rapporteur
Prof. Amaury Habrard	Université Jean Monnet	Examinateur
Dr. Georges Quénot	Équipe MRIM, LIG	Examinateur
Prof. Liming Chen	Ecole Centrale de Lyon	Directeur de thèse
Dr. Emmanuel Dellandréa	Ecole Centrale de Lyon	Co-encadrant de thèse

Abstract

In this dissertation we address the problem of weakly supervised object detection, wherein the goal is to recognize and localize objects in weakly-labeled images where object-level annotations are incomplete during training. To this end, we propose two methods which learn two different models for the objects of interest.

In our first method, we propose a model enhancing the weakly supervised Deformable Part-based Models (DPMs) by emphasizing the importance of location and size of the initial class-specific root filter. We first compute a candidate pool that represents the potential locations of the object as this root filter estimate, by exploring the generic objectness measurement (region proposals) to combine the most salient regions and "good" region proposals. We then propose learning of the latent class label of each candidate window as a binary classification problem, by training category-specific classifiers used to coarsely classify a candidate window into either a target object or a non-target class. Furthermore, we improve detection by incorporating the contextual information from image classification scores. Finally, we design a flexible enlarging-and-shrinking post-processing procedure to modify the DPMs outputs, which can effectively match the approximate object aspect ratios and further improve final accuracy.

Second, we investigate how knowledge about object similarities from both visual and semantic domains can be transferred to adapt an image classifier to an object detector in a semi-supervised setting on a large-scale database, where a subset of object categories are annotated with bounding boxes. We propose to transform deep Convolutional Neural Networks (CNN)-based image-level classifiers into object detectors by modeling the differences between the two on categories with both image-level and bounding box annotations, and transferring this information to convert classifiers to detectors for categories without bounding box annotations.

We have evaluated both our approaches extensively on several challenging de-

tection benchmarks, *e.g.* , PASCAL VOC, ImageNet ILSVRC and Microsoft COCO. Both our approaches compare favorably to the state-of-the-art and show significant improvement over several other recent weakly supervised detection methods.

Keywords: object detection, weakly supervised learning, deformable part models, region proposals, deep learning, convolutional neural networks, transfer learning

Résumé

Dans cette thèse, nous nous intéressons au problème de la détection d'objets faiblement supervisée. Le but est de reconnaître et de localiser des objets dans les images, n'ayant à notre disposition durant la phase d'apprentissage que des images partiellement annotées au niveau des objets. Pour cela, nous avons proposé deux méthodes basées sur des modèles différents.

Pour la première méthode, nous avons proposé une amélioration de l'approche "Deformable Part-based Models" (DPM) faiblement supervisée, en insistant sur l'importance de la position et de la taille du filtre racine initial spécifique à la classe. Tout d'abord, un ensemble de candidats est calculé, ceux-ci représentant les positions possibles de l'objet pour le filtre racine initial, en se basant sur une mesure générique d'objectness (par region proposals) pour combiner les régions les plus saillantes et potentiellement de bonne qualité. Ensuite, nous avons proposé l'apprentissage du label des classes latentes de chaque candidat comme un problème de classification binaire, en entrainant des classifieurs spécifiques pour chaque catégorie afin de prédire si les candidat sont potentiellement des objets cible ou non. De plus, nous avons amélioré la détection en incorporant l'information contextuelle à partir des scores de classification de l'image. Enfin, nous avons élaboré une procédure de post-traitement permettant d'élargir et de contracter les régions fournies par le DPM afin de les adapter efficacement à la taille de l'objet, augmentant ainsi la précision finale de la détection.

Pour la seconde approche, nous avons étudié dans quelle mesure l'information tirée des objets similaires d'un point de vue visuel et sémantique pouvait être utilisée pour transformer un classifieur d'images en détecteur d'objets d'une manière semi-supervisée sur un large ensemble de données, pour lequel seul un sous-ensemble des catégories d'objets est annoté avec des boîtes englobantes nécessaires pour l'apprentissage des détecteurs. Nous avons proposé de transformer des clas-

sifieurs d'images basés sur des réseaux convolutionnels profonds (Deep CNN) en détecteurs d'objets en modélisant les différences entre les deux en considérant des catégories disposant à la fois de l'annotation au niveau de l'image globale et l'annotation au niveau des boîtes englobantes. Cette information de différence est ensuite transférée aux catégories sans annotation au niveau des boîtes englobantes, permettant ainsi la conversion de classifieurs d'images en détecteurs d'objets.

Nos approches ont été évaluées sur plusieurs jeux de données tels que PASCAL VOC, ImageNet ILSVRC et Microsoft COCO. Ces expérimentations ont démontré que nos approches permettent d'obtenir des résultats comparables à ceux de l'état de l'art et qu'une amélioration significative a pu être obtenue par rapport à des méthodes récentes de détection d'objets faiblement supervisées.

Mots-clés: détection d'objets, apprentissage faiblement supervisé, deformable parts models, apprentissage profond, réseaux de neurones convolutionnels, transfert d'apprentissage

Acknowledgments

There are a number of people without whom this dissertation would never have been written. First and foremost, I would like to express my sincere and deep gratitude to my advisor, Prof. Liming Chen, for his academic guidance and enthusiastic encouragement throughout my PhD process. He showed me how to research a problem and achieve goals. He spent endless time reviewing and proofreading my papers, and supported me during the difficult times in my research. Besides, I am deeply grateful for precious advice and consistent support from my co-advisor Dr. Emmanuel Dellandréa, who has taught me a great deal. Working and discussing with him has truly strengthened my passion for science. The quality of this piece of work owes much to the creativity and insight from both my advisors.

I would like to thank the members of my PhD thesis committee Dr. Francesc Moreno-Noguer, Dr. Guoying Zhao, Prof. Amaury Habrard and Dr. Georges Quénot for accepting to evaluate this work and for their meticulous evaluations and valuable comments.

I give my thanks to all the collaborators (too many to list here) within the Visual Sense (Visen) project for our many discussions. A special thanks goes to Dr. Josiah Wang, for his thoughtful advise in the Natural Language Processing domain and for his help on the experiments, writing and proofreading of our CVPR paper. Many thanks to my co-authors Dr. Chao Zhu, Dr. Boyang Gao, Dr. Xiaofang Wang for their collaboration in this project.

In addition, I am also grateful to all the colleagues and friends in Lyon, including: Dr. Huanzhang Fu, Dr. Di Huang, Dr. Huibin Li, Dr. Dongming Chen, Dr. Huiliang Jin, Wuming Zhang, Li Wang, Wei Chen, Ying Lu, Huaxiong Ding, Yinhang Tang, Fei Zheng, Zehua Fu, Chen Wang, Hao Zhang, Xiangnan Yin and Haoyu Li, for their discussions and the happy time they have brought to me.

Moreover, many thanks to the Jacquier family and the Gachet family for their

enthusiastic help to make our life easier and happier in France.

I cannot end without thanking my family, my parents and my parents-in-law, for their constant care and encouragement. Specifically, this thesis is dedicated to Yang, my wife, for her unconditional love and company in the past years.

Contents

Abstract i

Résumé iii

Acknowledgments v

1 Introduction 1

1.1 Object Detection: Definition . . . 3

1.2 Challenges and Motivations . . . 4

1.2.1 General Challenges for Object Detection . . . 4

1.2.2 Challenges for Fully Supervised Object Detection . . . 5

1.2.3 Challenges for Weakly Supervised Object Detection . . . 7

1.2.4 Objective . . . 8

1.3 Approaches and Contributions . . . 9

1.3.1 Learning Weakly Supervised Deformable Part-Based Models for Object Detection Using Region Proposals . . . 9

1.3.2 Transferring Visual and Semantic Knowledge for Large Scale Semi-supervised Object Detection . . . 10

1.4 Outline . . . 11

2 Literature Review 13

2.1 Image and Object Representations . . . 14

2.1.1 Global Features . . . 15

2.1.2 Local Features . . . 15

2.1.2.1 Key points/regions detection . . . 16

2.1.2.2 Local descriptor extraction . . . 17

2.1.2.3 Feature encoding and aggregation . . . 18

2.1.3 Learned Features . . . 19

2.1.3.1 Supervised feature learning . . . 19

2.1.3.2 Unsupervised feature learning 19
2.2 Machine Learning Methods for Classification & Detection 20
2.2.1 Discriminative Approaches 20
2.2.2 Generative Approaches 21
2.2.3 Deep Learning 22
2.3 Fully Supervised Object Detection 24
2.3.1 Sliding Window Based Approaches 24
2.3.1.1 Deformable part-based models 25
2.3.2 Region Proposal Based Approaches 28
2.3.2.1 Region-based convolutional neural networks 30
2.4 Weakly Supervised Object Detection 35
2.4.1 Initialization Strategies 37
2.4.2 Iterative Learning Strategies 38
2.4.3 Transfer Learning Strategies 39

3 Weakly Supervised Learning of Deformable Part-Based Models for Object Detection via Region Proposals 43
3.1 Introduction 45
3.2 Fusing Generic Objectness and Deformable Part-Based Models for Weakly Supervised Object Detection 48
3.2.1 Object Estimations: Initialization 49
3.2.1.1 Region extraction 50
3.2.1.2 Salient reference region 50
3.2.1.3 Coarse candidate window pool 51
3.2.1.4 Object invariant estimations 51
3.2.2 Learning Latent Object Classes via Region Classification . . . 54
3.2.2.1 Region representation 55
3.2.2.2 Region classification 55
3.2.3 Weakly Supervised DPMs Training and Testing Details 57
3.2.3.1 Single region initialization for weak DPMs (S-WDPMs) detection 57

3.2.3.2 Multiple region initialization for weak DPMs (M-WDPMs) detection . . . 57
3.2.4 Bounding Box Post-processing . . . 59
3.3 Experimental Evaluation . . . 61
3.3.1 Experiments with S-WDPMs on PASCAL VOC Subsets . . . 62
3.3.1.1 Datasets and settings . . . 62
3.3.1.2 Evaluation protocol . . . 63
3.3.1.3 Experimental evaluation . . . 63
3.3.2 Experiments with M-WDPMs on PASCAL VOC . . . 66
3.3.2.1 Dataset and settings . . . 66
3.3.2.2 Parameter selection . . . 68
3.3.2.3 Annotation evaluation . . . 69
3.3.2.4 Detection evaluation . . . 72
3.3.2.5 Error analysis . . . 74
3.3.2.6 Running time . . . 76
3.3.3 Preliminary Results with M-WDPMs on MS COCO . . . 78
3.4 Summary . . . 79

4 Large Scale Semi-supervised Object Detection Using Visual and Semantic Knowledge Transfer 81
4.1 Introduction . . . 82
4.2 Task Definition . . . 86
4.3 Similarity-based Knowledge Transfer . . . 86
4.3.1 Background on LSDA . . . 86
4.3.2 Knowledge Transfer via Visual Similarity . . . 89
4.3.3 Knowledge Transfer via Semantic Relatedness . . . 91
4.3.4 Mixture Transfer Model . . . 93
4.3.5 Transfer on Bounding-box Regression . . . 94
4.4 Experiments . . . 96
4.4.1 Dataset Overview . . . 96
4.4.2 Implementation Details . . . 96

4.4.3 Quantitative Evaluation on the "Weakly Labeled" Categories with "*Alex-Net*" 97
4.4.4 Experimental Results with "*VGG-Nets*" 104
4.4.5 Experimental Results with Bounding-box Regression 105
4.4.6 Detection Error Analysis 105
4.5 Summary 108

5 Conclusion and Future Work **109**
5.1 Summary of Contributions 109
5.2 Perspective for Future Directions 111

6 List of Publications **115**

Bibliography **117**

List of Figures

1.1 An example of real world image that can be easily understood by a pre-school child. . . . 2
1.2 Object detection aims to recognize and localize objects of interest in images. . . . 3
1.3 General challenges in generic object detection. . . . 4
1.4 Some example images from PASCAL VOC 2007 with ground-truth bounding box annotations. . . . 6
1.5 A training example for weakly supervised object detection. . . . 7

2.1 Architecture of *LeNet-5* Convolutional Neural Networks (CNN) for handwritten digit recognition. . . . 23
2.2 Architecture of *AlexNet* Convolutional Neural Networks (CNN) for large-scale image classification. . . . 24
2.3 Detections obtained with a two component bicycle model of DPM. . . . 25
2.4 The matching process of DPM at one scale. . . . 27
2.5 An illustration of the Selective Search region proposal method. . . . 30
2.6 R-CNN object detector system overview. . . . 31
2.7 SPP-net network structure with a spatial pyramid pooling (SPP) layer. 32
2.8 Fast R-CNN object detector system overview. . . . 33
2.9 Faster R-CNN object detector system overview. . . . 34
2.10 An illustration of the Region Proposal Network (RPN) to generate region proposals. . . . 35
2.11 An illustration of Multiple Instance Learning (MIL) problem. . . . 36

3.1 Illustration of our proposed method to extract the initial object estimations. . . . 49
3.2 Some heat map examples generated by our methods. . . . 54
3.3 Illustration of our latent class learning framework for the *horse* category. . . . 56

3.4 Illustration of detection rescoring using an M-WDPMs and CNN softmax classifier. 59
3.5 Examples of bounding box enlarging-and-shrinking. 62
3.6 Examples of localization results for our S-WDPMs on PASCAL VOC 2007 images. 67
3.7 The impact of number of selected regions for each image in the multiple region initialization scheme. 69
3.8 Three example images and their 10 selected regions. 70
3.9 Analysis of top-ranked detections on PASCAL VOC 2007 test set. . . 75
3.10 Analysis of false positives for some classes on which our M-WDPMs outperforms supervised DPMs. 76
3.11 Detection results of weakly supervised DPMs detectors on MS COCO 2014 val2. 77

4.1 Semi-supervised object detection: Problem definition. 83
4.2 An illustration of our similarity-based knowledge transfer model. . . 84
4.3 The pipeline of the LSDA framework. 88
4.4 Some example visualizations of visual, semantic and mixture similarity between a target category and its source categories. 99
4.5 Sensitivity of parameter α vs. mAP. 101
4.6 Examples of correct detections of our mixture knowledge transfer model on ILSVRC2013 images. 102
4.7 Examples of incorrect detections of our mixture knowledge transfer model on ILSVRC2013 images. 103
4.8 Some example detections before and after bounding box regression on the "weakly labeled" categories. 106
4.9 Analysis of detection errors of our model. Error trend and fractions before and after bounding-box regression are compared. 107

List of Tables

3.1 Average localization accuracy of our S-WDPMs compared with state-of-the-art competitors on the two variations of the PASCAL VOC 2007 datasets. 64

3.2 Comparison of class level localization accuracy for the *VOC07-6×2* dataset. 66

3.3 Comparison of weakly supervised object detectors on PASCAL VOC 2007 trainval set in terms of correct localization. 71

3.4 Comparison of weakly supervised object detectors on PASCAL VOC 2007 in terms of AP in the test set. 73

4.1 Mean average precision (mAP) of detection results on ILSVRC2013 val2 dataset. 98

4.2 Comparison of mean average precision (mAP) for semantic similarity measures/representations, using **Weighted - 100**. 101

4.3 Comparison of detection mean average precision (mAP) on the "weakly labeled" categories of ILSVRC2013 val2, using the "***VGG-Nets***". For LSDA, our visual similarity and semantic relatedness transfer models, **Weighted - 100** scheme is adopted. 105

Chapter 1

Introduction

Contents

1.1 Object Detection: Definition 3

1.2 Challenges and Motivations 4

1.2.1 General Challenges for Object Detection 4

1.2.2 Challenges for Fully Supervised Object Detection 5

1.2.3 Challenges for Weakly Supervised Object Detection 7

1.2.4 Objective 8

1.3 Approaches and Contributions 9

1.3.1 Learning Weakly Supervised Deformable Part-Based Models for Object Detection Using Region Proposals 9

1.3.2 Transferring Visual and Semantic Knowledge for Large Scale Semi-supervised Object Detection 10

1.4 Outline 11

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc.

-Alan Turing, *Computing Machinery and Intelligence (1950)*

Figure 1.1: An example of real world image that can be easily understood by a pre-school child. Image from Microsoft COCO dataset [Lin *et al.* 2014b].

According to Alan Turing [Turing 1950], the goal of Artificial Intelligence is to create intelligent machines that can think and act humanly. To achieve this goal, the intelligent machines should be able to "function appropriately and with foresight in their environment" [Nilsson 2009] like humans do. Visual perception is the ability to interpret the surrounding environment by processing information that is contained in visible light. Humans are extraordinarily capable of perceiving and understanding the rich visual world. Even a pre-school child can easily recognize the objects (*e.g.* , person, bus, car, traffic light) in real world images as Figure 1.1. While in a computer, this image is represented as an array of numbers indicating the brightness at any position.

Over the last few decades, an explosive growth of image and video data digitally available both online and offline (*e.g.* , social media sharing websites/applications and personal photo albums) drives the computer vision community to make endeavors to empower machines to mimic humans' ability to discover and understand the visual content. A significant aspect of image content is

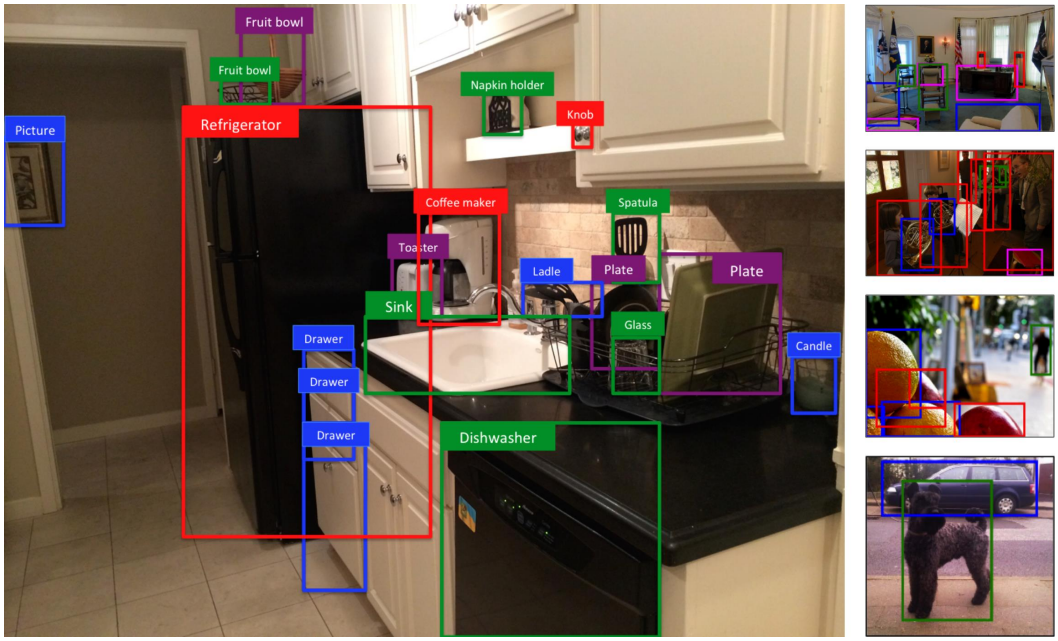

Figure 1.2: Object detection aims to recognize and localize objects of interest in images. Image courtesy of [Russakovsky 2015].

the object composition: the identities and positions of the objects the images contain (See Figure 1.2). Designing an intelligent visual understanding model that can recognize the object composition gives rise to many challenging applications such as automatic image annotation, image retrieval, self-driving cars, robotics, video surveillance, searching online shopping catalogs, home and health-care automation, etc.

1.1 Object Detection: Definition

Visual (scene) understanding started with the goal of building machines that can see like humans to infer general principles and current situations from imagery. Typically, the most essential component of an intelligent visual understanding model is its object detection module [Andreopoulos & Tsotsos 2013]. We begin with a definition of the object detection problem. Given an arbitrary image, an ideal object detection method aims not only to recognize but also to locate objects in categories of interest within the image. Figure 1.2 shows an object detection ex-

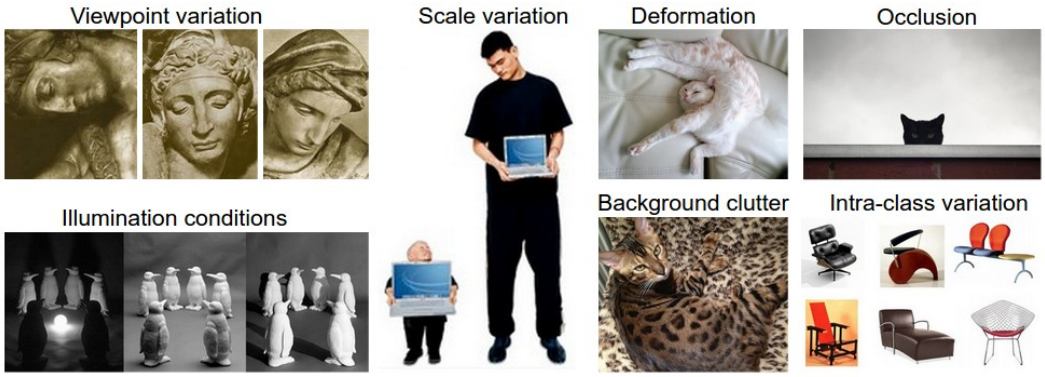

Figure 1.3: General challenges in generic object detection. Image courtesy: http://cs231n.github.io/classification/

ample that recognize all the objects of interest and determine the location together with the extent of each object instance by drawing a rectangular bounding box.

The success of early detection methods starts from localizing constrained object categories, such as face [Viola & Jones 2001] or pedestrian [Dalal & Triggs 2005]. Recent approaches are moving to the detection of various categories of generic objects with large appearance variations, *e.g.* , from the 20 categories of PASCAL VOC [Everingham *et al.* 2010] to 80 categories of Microsoft COCO [Lin *et al.* 2014b] and 200 categories of ImageNet ILSVRC [Russakovsky *et al.* 2015]. In this dissertation, we focus on object detection of generic object categories, since it is a necessary first step to the scene understanding problem.

1.2 Challenges and Motivations

1.2.1 General Challenges for Object Detection

There are many difficulties and challenges associated with the object detection task, most of which are caused by the fact that visual appearances of objects in images vary largely by the following factors:

Viewpoint variation: An important cause of the variability in visual appearance is that a single instance of an object can be oriented in many ways with respect to the camera during image acquisition.

Illumination conditions: A variation in illumination condition can change the

pixel intensities of an object surface. A good object detection system should be invariant to several illumination transformations.

Scale variation: Variation in size of same object category often exists not only in terms of their extent in the image, but also in the real world. An ideal detector should be able to detect object instances existing in various scales.

Deformation: Many objects of interest are not rigid bodies thus can undergo various deformations. For example, animals and people can have different poses and actions such as walking, sitting, lying, etc. A successful object detection method must have the ability to recognize objects across a wide range of deformations.

Background clutter: In real world images, the objects of interest may blend into their environment, making them hard to identify.

Occlusion: Some parts of objects of interest can be occluded by other objects or stuff. Sometimes only a small portion of an object could be visible due to occlusion.

Intra-class variation: Within an object category, the object instances can vary significantly in forms of color, texture, shape, number of parts, etc. For example, for the *aeroplane* class from the PASCAL VOC dataset, fighter jet, jumbo passenger jet, single-engine propeller plane, and biplane are all labeled as "*aeroplane*", although they are visually very dissimilar.

Figure 1.3 shows some examples of different general challenges in generic object detection.

In addition to these variations, the similarity between object categories also exists in real world images. For example, a small number of object instances annotated as "dog" and "cat" are visually very similar, which are fairly hard to differentiate.

1.2.2 Challenges for Fully Supervised Object Detection

For most of the object detection methods, a fully supervised learning (FSL) approach is adopted [Dalal & Triggs 2005, Felzenszwalb *et al.* 2010b, Szegedy *et al.* 2013, Girshick *et al.* 2014], where positive training images are manually annotated with bounding boxes encompassing the objects of inter-

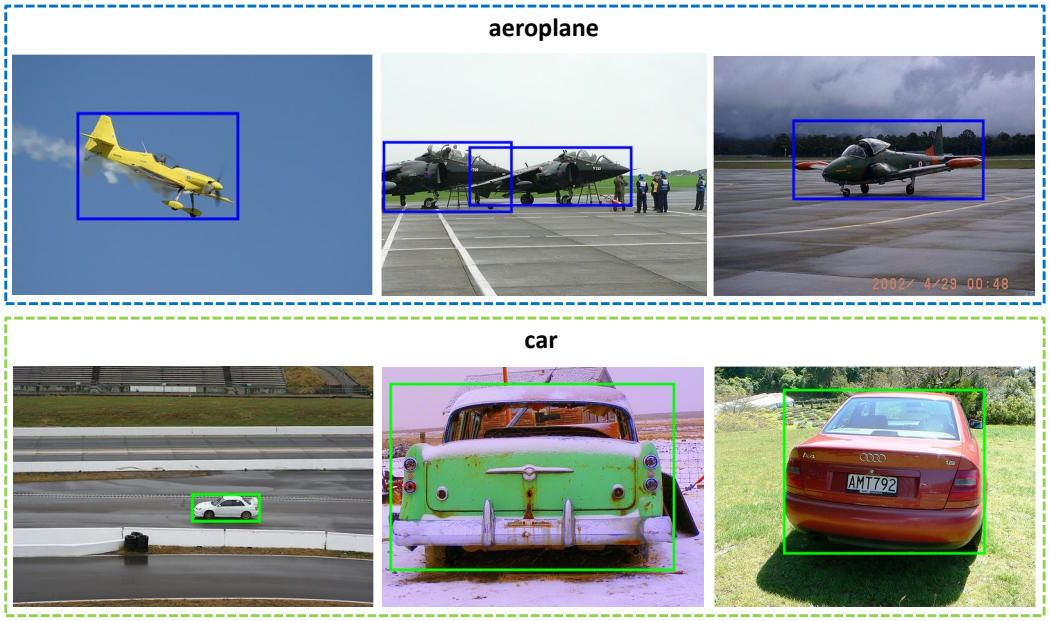

Figure 1.4: Some example images from PASCAL VOC 2007 with ground-truth bounding box annotations.

est. We show some examples of PASCAL VOC 2007 [Everingham *et al.* 2010] bounding box annotations in Figure 1.4. Most conventional fully supervised methods for object detection learn a discriminative model on these fully annotated data. During training, they typically reduce a detection problem to a binary classification problem by treating different object categories independently. Binary classifiers run within the image in a sliding window manner [Dalal & Triggs 2005, Felzenszwalb *et al.* 2010b], or on a set of region proposals [Alexe *et al.* 2012, Uijlings *et al.* 2013] to considerably reduce computation effort.

Fully supervised learning methods has been improving rapidly during the last decade in terms of both accuracy and speed [Dalal & Triggs 2005, Felzenszwalb *et al.* 2010b, Girshick *et al.* 2014, Girshick 2015, Ren *et al.* 2015a, Liu *et al.* 2016]. Moreover, the key to the success of deep Convolutional Neural Networks (CNN) [Krizhevsky *et al.* 2012] for object detection is the ability to learn from large quantities of fully annotated data. However, fully supervised learning is not scalable due to the lack of fully annotated data, especially for large scale data. For example, only about 3,000 of 21,841 synsets in ImageNet dataset

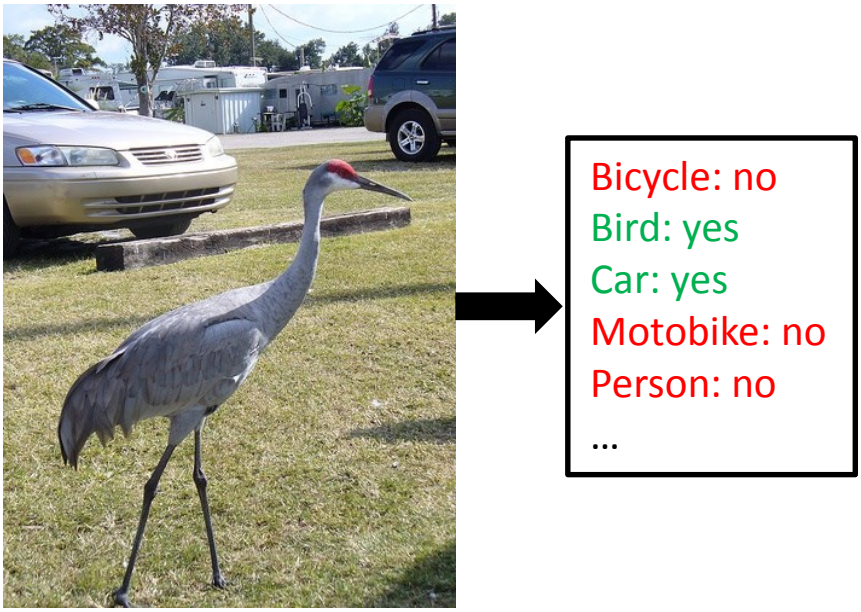

Figure 1.5: A training example for weakly supervised object detection. Image from PASCAL VOC 2007 [Everingham *et al.* 2010].

[Russakovsky *et al.* 2015] are annotated with bounding boxes. Therefore, fully supervised object detection has limitations in extending to new object categories without manually labeled bounding box annotations.

1.2.3 Challenges for Weakly Supervised Object Detection

Although annotations on objects localization are extremely valuable, the process of manually annotating object bounding boxes is extremely laborious and unreliable, especially for large-scale databases. For example, annotating the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset [Russakovsky *et al.* 2015] for object detection, which has hundreds of thousands of images of 200 object categories required 42 seconds per bounding-box by crowd-sourcing on Amazon Mechanical Turk[1]. On the other hand, it is usually much easier to obtain annotations at image level, indicating the absence or presence of object instances of the category (A training example is shown in Figure 1.5). For example, from user-generated tags on Flickr or Web queries. One could directly apply classifiers trained at image-level

[1] https://www.mturk.com/

to detect object categories, but it performs poorly as there are differences in the statistical distribution between the training data (whole images) and the test data (localized object instances). This phenomenon is known as "domain shift".

In recent years, there has been a substantial amount of work on weakly supervised object detection. Based on weakly annotated examples, the common practice is to jointly learn an appearance model together with the latent object location. The majority of related work treats the weakly supervised object detection as a multiple instance learning (MIL) [Maron & Ratan 1998] problem. In the MIL framework, there are some positive and some negative bags. A bag is positive when it has at least one positive instance, while it is negative if all the instances are negative. The object detector is then obtained by alternating detector training, and using the detector to select the most likely object instances in positive images. In most MIL framework, there is a huge number of examples in each bag when using exhaustive searching (*e.g.* , sliding window approach). Although region proposal methods [Alexe *et al.* 2012, Uijlings *et al.* 2013, Zitnick & Dollar 2014, Bilen *et al.* 2014] can significantly reduce the search space per image, the selection of windows across a large number of images is inherently a challenging problem, where an iterative weakly supervised method can typically find only a local optimum depending on the initial windows [Cinbis *et al.* 2014]. A good initialization is crucial, however, this is a "chicken and egg" problem: a good model can be learned if we start from a selection of good initial windows, but the latent annotations are needed to infer a good model. Moreover, multiple objects and small objects are widely existed in recent dataset such as PASCAL VOC [Everingham *et al.* 2010], ImageNet [Russakovsky *et al.* 2015], Microsoft COCO [Lin *et al.* 2014b], which makes the task of weakly supervised object detection more challenging.

1.2.4 Objective

Based on the above discussion, in this dissertation, in contrast to the traditional fully supervised learning (FSL), we are concerned with weakly supervised learning (WSL) for object detection, where image-level labels indicating the presence or the

absence of the object are given, but the exact object locations in positive training examples are not provided or only partially provided.

Deformable Part-based Models (DPMs) [Felzenszwalb *et al.* 2010b] and Region-based Convolutional Neural Networks CNNs [Girshick *et al.* 2016, Girshick 2015, Ren *et al.* 2015a] are successful frameworks for object detection in recent years. Therefore, our objective is to propose approaches with weak supervision based on these two models.

1.3 Approaches and Contributions

In this dissertation, we propose two novel approaches for weakly supervised object detection. Our approaches and contributions are summarized in the following subsections.

1.3.1 Learning Weakly Supervised Deformable Part-Based Models for Object Detection Using Region Proposals

We propose a model enhancing the weakly supervised Deformable Part-based Models (DPMs) by emphasizing the importance of location and size of the initial class-specific root filter. To adaptively select a discriminative set of candidate bounding boxes as this root filter estimate, first, we explore the generic objectness measurement (region proposals) to combine the most salient regions and "good" region proposals. Second, we propose learning of the latent class label of each candidate window as a binary classification problem, by training category-specific classifiers used to coarsely classify a candidate window into either a target object or a non-target class. Furthermore, we incorporate the contextual information from image classification, by combining the image-level classification score with object-level DPM detection score, to obtain a final score for detection so as to improve detection. Finally, we design a flexible enlarging-and-shrinking post-processing procedure to modify the DPMs outputs, which can effectively match the approximate object aspect ratios and further improve final accuracy. Extensive experimental results on the challenging PASCAL Visual Object Class (VOC) 2007 and the

Microsoft Common Objects in Context (MS COCO) 2014 dataset demonstrate that our proposed framework is effective for initialization of the DPMs root filter. It also shows competitive final localization performance with state-of-the-art weakly supervised object detection methods, particularly for the object categories which are relatively salient in the images and deformable in structures. This work is published and was awarded the Top 10% Paper Award at the IEEE International Conference on Image Processing (ICIP) 2014 [Tang *et al.* 2014b] and its extended version [Tang *et al.* 2016b] is published in the IEEE Transactions on Multimedia.

1.3.2 Transferring Visual and Semantic Knowledge for Large Scale Semi-supervised Object Detection

We investigate how knowledge about object similarities from both visual and semantic domains can be transferred to adapt an image classifier to an object detector in a semi-supervised setting, where a subset of object categories are annotated with bounding boxes. We propose to transform deep CNN-based image-level classifiers into object detectors by modeling the differences between the two on categories with both image-level and bounding box annotations, and transferring this information to convert classifiers to detectors for categories without bounding box annotations. The intuition behind our proposed method is that visually and semantically similar categories should exhibit more common transferable properties than dissimilar categories, *e.g.* , a better detector would result by transforming the differences between a *dog* classifier and a *dog* detector onto the *cat* class, than would by transforming from the *violin* class. Experimental results on the challenging ILSVRC2013 detection dataset demonstrate that each of our proposed object similarity based knowledge transfer methods outperforms the baseline methods. We found strong evidence that visual similarity and semantic relatedness are complementary for the task, and when combined notably improve detection, achieving state-of-the-art detection performance in a semi-supervised setting. This work is published in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016 [Tang *et al.* 2016a] and its extended version will be submitted to a top

journal.

1.4 Outline

The remaining of this dissertation is organized as follows:

In **Chapter 2** we introduce the fundamental of object detection and reviews the literature of various fully supervised and weakly supervised object detection methods.

In **Chapter 3** we present our selection model to learn Deformable Part-based Models in a weakly supervised manner.

In **Chapter 4** we present our knowledge transfer method to transform a CNN image classifier into an object detector in a semi-supervised manner.

In **Chapter 5** we conclude our work and discuss the remaining challenges.

Finally, in **Chapter 6** we list our publications.

Chapter 2

Literature Review

Contents

2.1 Image and Object Representations 14
2.1.1 Global Features 15
2.1.2 Local Features 15
2.1.2.1 Key points/regions detection 16
2.1.2.2 Local descriptor extraction 17
2.1.2.3 Feature encoding and aggregation 18
2.1.3 Learned Features 19
2.1.3.1 Supervised feature learning 19
2.1.3.2 Unsupervised feature learning 19
2.2 Machine Learning Methods for Classification & Detection 20
2.2.1 Discriminative Approaches 20
2.2.2 Generative Approaches 21
2.2.3 Deep Learning 22
2.3 Fully Supervised Object Detection 24
2.3.1 Sliding Window Based Approaches 24
2.3.1.1 Deformable part-based models 25
2.3.2 Region Proposal Based Approaches 28
2.3.2.1 Region-based convolutional neural networks 30
2.4 Weakly Supervised Object Detection 35
2.4.1 Initialization Strategies 37
2.4.2 Iterative Learning Strategies 38
2.4.3 Transfer Learning Strategies 39

Object detection involves locating the target objects of particular categories in the image. Generally, an object detector can be regarded as a combination of an image feature set and a detection algorithm. Image feature representation involves dense or sparse extraction of features vectors from image patches; and the detection architecture engages learning algorithms to recognize instances of an object category based on the feature vectors.

In this chapter, we provide an overall review of the literature on object detection related work. We start by briefly reviewing image and object representations in Section 2.1. Then, we introduce some general classification and detection methods in Section 2.2. Next, we provide an overview of the fully supervised object detection approaches in Section 2.3. Finally, we review the existing weakly supervised learning models for object detection, where the goal is to learn object detection models with no or incomplete bounding box annotations.

In our overview, we focus mainly on the close related work of this dissertation, while a detailed review of the object recognition methods can be found in the excellent survey paper by [Andreopoulos & Tsotsos 2013].

2.1 Image and Object Representations

The first step of object recognition is to transform the content of an image or an image region into a set of feature vectors, or descriptors, which are expected to discriminatively represent the image content with efficient computation, reasonable size, and robustness to image variations resulted by illumination, scale, pose, etc.. The feature exaction step is critical to ensure good detection performance, and is considered as the basis of the whole detection process.

Extensive literature exists on the exploitation of feature extraction, in order to improve recognition. Prior to deep Convolutional Neural Network (CNN) feature [Krizhevsky *et al.* 2012], there were two contrasting views in computer vision on how to compute feature vectors: global features and local features.

2.1.1 Global Features

Early research on appearance-based recognition mainly used global features based on color, texture or shape histograms from the whole image.

Color is arguably the most direct and expressive visual information. Methods for extracting color features (*e.g.* , color histogram [Swain & Ballard 1991], color moments [Stricker & Orengo 1995], color coherence vectors [Pass *et al.* 1996]) capture color information, such as color distribution, relationship between different colors, etc., contained in an image.

Texture features can be intuitively considered as the repeated patterns of local variation of pixel intensities. Gabor [Daugman 1988] filters are widely adopted to extract global texture features for image analysis.

Shape is a geometrical description of the external boundary of an object which can be described by basic geometry units such as points, lines, curves and planes. Popular shape features [Park *et al.* 2000, Pujol & Chen 2007] mainly focus on the edge or contour of an object to capture its shape information.

The raw extracted global features can be projected into a lower dimensional feature space, for instance, using Principal Component Analysis (PCA) [Murase & Nayar 1995], which is more easily amenable to powerful classifiers for recognition.

The main drawback of aforementioned global features is that they are very sensitive to background clutter, occlusion, and illumination variations. Moreover, these global methods implicitly assume that objects of interest take up most of the image content, which is not always desirable in real world images. All these constraints make global features gradually give their way to local features.

2.1.2 Local Features

Local feature based recognition methods have drawn a lot of attention since the second half of the 1990s due to their robustness in background clutter and partial occlusion. Almost twenty years passed, local features are still at the center of many fundamental computer vision problems in-

cluding registration, stereo vision, motion estimation, matching, retrieval, recognition of objects and actions. Local features can be based on points [Harris & Stephens 1988, Mikolajczyk & Schmid 2002], blobs (Laplacian of Gaussian [Lindeberg 1998] or Difference of Gaussian [Lowe 2001]), intensities [Kadir & Brady 2001], color [Zhu *et al.* 2013], texture (Local Binary Patterns (LBP) [Ojala *et al.* 2002] and its variants which have extensive applications in texture classification [Zhao *et al.* 2012], face recognition [Ahonen *et al.* 2006], facial expression detection [Zhao & Pietikainen 2007], etc.), gradient [Mikolajczyk *et al.* 2004], or combinations of several of these. Generally, local feature extraction consists of three main steps: (1) key points/regions detection, (2) local descriptor extraction from the detected key points/regions, (3) feature encoding and aggregation.

2.1.2.1 Key points/regions detection

Local features are extracted from local image regions, thus it is important to obtain a representative set of image regions/patches covering the essential information of a given image. There are two mainly strategies for this purpose. The first one is sparse sampling based on points, image fragments or part detectors and the second one is dense sampling using image intensities or gradients.

Sparse sampling: Interest point detectors aim to find a sparse set of discriminative regions containing plenty of information about image structures like edges and corners, or local blobs with uniform brightness. The final detector are then based on the feature vectors computed from the extracted key points/regions. Many key point/region detectors have been proposed in the literature, the most commonly used key point/region detectors include Harris [Harris & Stephens 1988], Laplacian [Lindeberg 1998], Difference of Gaussians (DoGs) [Lowe 2004], scale invariant Harris-Laplace [Mikolajczyk & Schmid 2004], maximally stable extremal regions (MSER) [Matas *et al.* 2004]. An advantage of sparse sampling by finding key points is that the number of interest points is much fewer than that of image pixels, thus generating a compact representation, which speeds up latter classification/detection process. However, most of the key point detectors

are designed to fire repeatedly on particular objects and might have constraints when generalizing to generic object categories. Comprehensive reviews and evaluations of key point/region detectors can be found in [Schmid *et al.* 2000] and [Mikolajczyk *et al.* 2005].

Dense sampling: Another approach is to extract visual features densely (often pixel-wise) over an entire image or detection window and to collect them into a high-dimensional descriptor vector that can be used for discriminative image classification or object recognition. Typically the representation is based on image intensities [Vidal-Naquet & Ullman 2003], gradients [Ronfard *et al.* 2002] or higher order differential operators [Viola & Jones 2001]. Dense sampling methods have been observed to outperform the sparse sampling methods for object recognition [Nowak *et al.* 2006]. One potential reason for this is that dense sampling avoids losing significant information by sampling uniformly from the entire image, while sparse sampling may skip some crucial information for object recognition by only looking around key points/regions. In order to have decent categorization accuracy, the sampling frequency should be increased to ensure obtaining similar patches across images, which results in a significant increase in feature extraction cost. [Tuytelaars 2010] proposes to find interest points with a dense grid to alleviate this issue.

2.1.2.2 Local descriptor extraction

After sampling image patches, the next step is to extract feature vectors (or local descriptors). Various of local descriptors have been proposed in the literature, and the most popular ones are distribution-based descriptors, which represent region properties by histograms. The most popular local descriptors applied to object recognition include: Scale Invariant Feature Transform (SIFT) [Lowe 2004], Histogram of Oriented Gradient (HOG) [Dalal & Triggs 2005], DAISY [Winder *et al.* 2009], Speeded-Up Robust Features (SURF) [Bay *et al.* 2008] and shape context [Belongie *et al.* 2002]. There have been numerous research studies contributed to improve the pioneer work of [Lowe 2004, Dalal & Triggs 2005]

by integrating new features. The combination of these feature with color or motion information has proved to increase the detectors performance. These local descriptors are designed to be discriminative, computationally efficient, and robust against various image variations such as scaling, affine distortions, viewpoint and illumination variations.

2.1.2.3 Feature encoding and aggregation

After local feature extraction, each image is represented by a set of local descriptors. Efficient feature encoding and aggregation methods are required to transform the high dimensional raw local descriptors into a more compact, informative and fixed-length representation for constructing a suitable image descriptor.

Bag-of-words (BoW, a.k.a. bag-of-visual-words (BoVW), bag-of-features(BoF)) [Sivic & Zisserman 2003, Csurka *et al.* 2004] has been one of the most popular image representation methods. The key idea of BoW is to represent an image with orderless distributions of local image features based on an intermediate representation called visual vocabulary. Typically it consists of three steps: (1) visual vocabulary (a.k.a. dictionary or codebook) construction (*e.g.* , using k-means clustering [Macqueen 1967], Guassian mixture models (GMM) [Fernando *et al.* 2012]), (2) feature encoding (*e.g.* , sparse coding [Olshausen & Field 1997]), Super Vector (SV) [Zhou *et al.* 2010], Fisher Vector (FV) [Sánchez *et al.* 2013]), (3) feature aggregation (*e.g.* , average [Carreira *et al.* 2012], max pooling [Boureau *et al.* 2010]). A comprehensive analysis and evaluation of feature encoding and aggregation methods can be found in [Chatfield *et al.* 2011].

Since BoW method views images as an orderless collection of local features, the spatial relationships between them were not explicitly considered. The most popular method for incorporating global layout into the image representation is the Spatial Pyramid Matching (SPM) proposed by [Lazebnik *et al.* 2006]. This provides a mid-level representation which bridges the semantic gap between low-level features extracted from an image patch and high-level concepts to be categorized.

2.1.3 Learned Features

The aforementioned global and local features are hand-crafted features which are usually designed for specific purposes: HOG for appearance and shape, LBP for texture, shape context for shape, etc., and their modeling capacities are limited by the fixed transformations (filters) that stay the same for different sources of data. On the other hand, feature learning, which aims to automatically learn useful features or representations from raw data without requiring expensive human labor or expert knowledge, has received a lot of attention. Typically, feature learning methods can be categorized into two different groups: (1) supervised and (2) unsupervised learning.

2.1.3.1 Supervised feature learning

In supervised feature learning, features are learned with labeled input data. Examples include neural networks [Bishop 1995], multilayer perceptron (MLP) [Attardi *et al.* 2009], and supervised dictionary learning [Mairal *et al.* 2009]. In the last few years, a prominent supervised feature learning example is Convolutional Neural Networks (CNN or ConvNet) [LeCun *et al.* 1990, Lecun *et al.* 1998, Krizhevsky *et al.* 2012]. Features extracted from the intermediate layers of a CNN are proved to be very powerful for many computer vision tasks. [Simo-Serra *et al.* 2015] learn compact discriminative feature point descriptors using a CNN to represent an image patch. We will elaborate the architecture of CNNs in Section 2.2.3.

2.1.3.2 Unsupervised feature learning

In unsupervised feature learning, features are learned with unlabeled input data. Examples include dictionary learning [Lee *et al.* 2006], independent component analysis (ICA) [Hyvärinen & Oja 2000], sparse autoencoders [Makhzani & Frey 2014], matrix factorization [Srebro *et al.* 2005], and various forms of clustering [Csurka *et al.* 2004]. An autoencoder neural network is an unsupervised learning algorithm that applies backpropagation, setting the target values

to be equal to the inputs by learning a function $h_{W,b}(x) \approx x$. Sparse coding is a class of unsupervised methods for learning sets of over-complete bases to represent data efficiently. The aim of sparse coding is to find a set of basis vectors $\phi(i)$ such that we can represent an input vector **x** as a linear combination of these basis vectors: $\mathbf{x} = \sum_{i=1}^{k} a(i)\phi(i)$.

2.2 Machine Learning Methods for Classification & Detection

Object detection systems construct a model for each object category from a set of training examples. Generally, object detection methods can be categorized into discriminative approaches and generative approaches [Amit & Felzenszwalb 2014]. Both discriminative and generative models start with an initial choice of image features. The principal differences between discriminative and generative models are in the methods of training and computation. A significant distinction is that discriminative methods need data from both foreground object and background regions to learn the decision boundaries whereas generative models do not need data from background to train the object models. In this section, we will review the popular classification methods for object recognition.

2.2.1 Discriminative Approaches

Discriminative approaches typically build a classifier that can discriminate between images (or image regions) containing the object instance and those not containing the object. The parameters of the classifier are learned to minimize errors on the training data, often with a regularization term to avoid overfitting. Machine learning techniques such as Support Vector Machine (SVM) [Cortes & Vapnik 1995] and Boosting [Schapire 2001] have become popular as classifiers for object recognition owing to their ability to automatically select relevant descriptors or features from large feature sets and their decent performance.

Support Vector Machine (SVM) classifiers have been widely used for object

recognition for the past decade [Cortes & Vapnik 1995]. SVM constructs a hyperplane in a high or infinite dimensional space to separate the samples from different classes for classification. A good separation is is to construct a decision boundary (hyperplane) that maximizes the margin (gap) between the object class and non-object class in either the input feature space or a kernelized version of this. The maximum margin problem can be formulated as the following unconstrained convex optimization problem:

$$\arg\min_{w} J(w) = \arg\min_{w}\{C\sum_{i}\max(0, 1 - y_i(w^T x_i + b)) + \frac{1}{2}\|w\|^2\} \tag{2.1}$$

where $J(w)$ is the objective function, (w, b) is the hyperplane, x is a sample, $y \in \{+1, -1\}$ is the label and C is a trade-off parameter that penalizes the margin violations.

Adaptive Boosting (a.k.a. AdaBoost) [Freund & Schapire 1997] gathers a collection of weak classifiers to form a stronger one, which is used particularly to build cascades of pattern rejecters, with at each level of the cascade choosing the features that are most relevant for its rejection task. Although AdaBoost cascades take relatively long time to train, owing to their selective feature encoding they offer significant improvement (compared to SVMs) in the run-time of the final detectors.

2.2.2 Generative Approaches

In contrast to discriminative approaches, generative approaches produce a probability density (joint probability distribution) model over all the variables and then adopt it to compute classication functions. The most common generative models are based on Bayesian or graphical models. [Weber *et al.* 2000] adopt Bayesian generative models learned with expectation maximization (EM) to characterize classes, and use likelihood ratios for classification. [Fergus *et al.* 2003] also use likelihood ratios, but with a more elaborate model of conditional probabilities that includes the position and scale of the features as well as their appearance.

2.2.3 Deep Learning

In recent years, deep learning based methods such as deep Convolutional Neural Networks (CNN) have become prominent since their ability to learn concepts with minimal feature engineering, end-to-end and in a purely data driven fashion. Typically, a CNN is composed of a sequence of layers that convolve the input image with filters, apply non-linear transformations on filter responses and spatially pool the resulting values. CNNs are very similar to ordinary neural networks (NN) since they are both consist of neurons that have learnable weights and biases, but they are different from NNs: (1) CNNs use convolution in place of general matrix multiplication in at least one of their layers, (2) the number of parameters in CNNs is significantly reduced in comparison with fully connected NNs due to weights sharing. Although it has been decades since the introduction of CNNs [LeCun *et al.* 1990, Lecun *et al.* 1998] (best known work is *LeNet-5* for handwritten digit recognition, see Figure 2.1), only recently CNNs have seen a surge of attention from the computer vision community [Krizhevsky *et al.* 2012], demonstrating previously unattainable performance on the tasks of image classification[Krizhevsky *et al.* 2012, He *et al.* 2016], object detection [Girshick *et al.* 2016, Girshick 2015, Ren *et al.* 2015a], localization [Sermanet *et al.* 2014] and semantic segmentation [Girshick *et al.* 2014]. The improvements in CNN training techniques(*e.g.* , rectified linear unit (ReLU) non-linearity [Nair & Hinton 2010]), computational resources (*e.g.* , large memory GPUs) and large scale datasets (*e.g.* , ImageNet [Russakovsky *et al.* 2015]) have made the CNN-based architectures more powerful.

In CNNs, a feature map can be obtained after a series of convolutional operation followed by an element-wise non-linearity. Each convolutional layer is typically followed by a pooling layer – an operation which outputs a maximum (or an average, *i.e.* max pooling or average pooling) value within a local neighborhood, reducing the size of a feature map and introducing additional invariance to small local translations. An example of first major success of deep learning in computer vision on large scale image dataset for image classification is the *AlexNet*

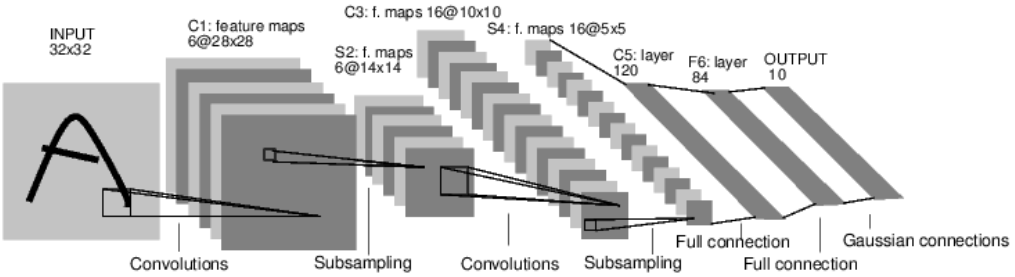

Figure 2.1: Architecture of *LeNet-5* Convolutional Neural Networks (CNN) for handwritten digit recognition.

CNN [Krizhevsky *et al.* 2012]. The architecture of *AlexNet* is shown in Figure. 2.2. It has eight layers (five convolutional layers and three fully connected layers) and it contains 60 million parameters, which are automatically learned on the ImageNet dataset containing 1.2 million training images of 1000 image categories. The *AlexNet* was the winner of the ImageNet ILSVRC challenge for image classification in 2012 and significantly outperformed the second runner-up (top 5 error of 16% compared to runner-up with 26% error). A number of more advanced and complex convolutional networks, especially in the context of visual recognition and detection from images, have been proposed over last years, including Zeiler & Fergus Net (*ZF-Net*) [Zeiler & Fergus 2014], Network in Network (*NIN*) [Lin *et al.* 2014a], Inception modules (or *GoogLeNet*) [Szegedy *et al.* 2015], VGG-Nets (*VGG-16* and *VGG-19*) [Simonyan & Zisserman 2015]. Recently, the Residual network (*ResNet*) [He *et al.* 2016] with more than 150 layers won the first place for many tasks in ILSVRC [Russakovsky *et al.* 2015] and COCO [Lin *et al.* 2014b] challenge in 2015.

Since a deep learning model implicitly learns an image representation, the upper-layers of a pre-trained model on a large scale dataset can be used as a high-level feature extractor. For instance, [Girshick *et al.* 2014] use $fc6$ and $fc7$ layer output of the *AlexNet*, which is a 4096-dimensional feature vector, as an input to SVM for classification.

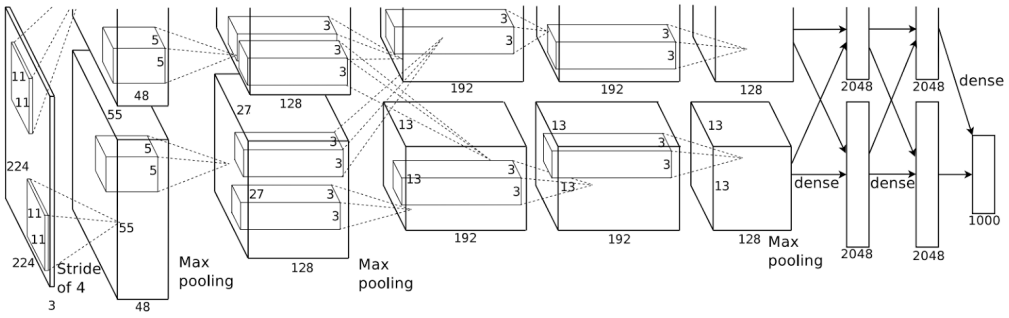

Figure 2.2: Architecture of *AlexNet* [Krizhevsky *et al.* 2012] Convolutional Neural Networks (CNN) for large-scale image classification.

2.3 Fully Supervised Object Detection

The most common approach to object detection reduces the problem to a set of binary classification problems by applying class-specific classifiers which classifying each candidate window (sub-image) into target category that contains the object of interest, or non-target category that does not contain the object. In this section, we review the fully supervised object detection [Dalal & Triggs 2005, Felzenszwalb *et al.* 2010b, Szegedy *et al.* 2013, Girshick *et al.* 2014, Ren *et al.* 2015a, Liu *et al.* 2016], where positive training images are manually annotated with bounding boxes encompassing the objects of interest. Localization Strategies can be divided into two different groups: (1) sliding window approaches and (2) region proposal approaches.

2.3.1 Sliding Window Based Approaches

In the context of computer vision, a sliding window is rectangular region of fixed width and height that "slides" across an image. The sliding window approach to object detection involves explicitly considering and classifying every possible window over an exhaustive list of positions, scales, and aspect ratios. For each of these windows, some class-specific image classifiers can be applied to it, to determine if the window has an object of interest. Normally, utilizing both a sliding window and an image pyramid we are able to detect objects in images at various scales and

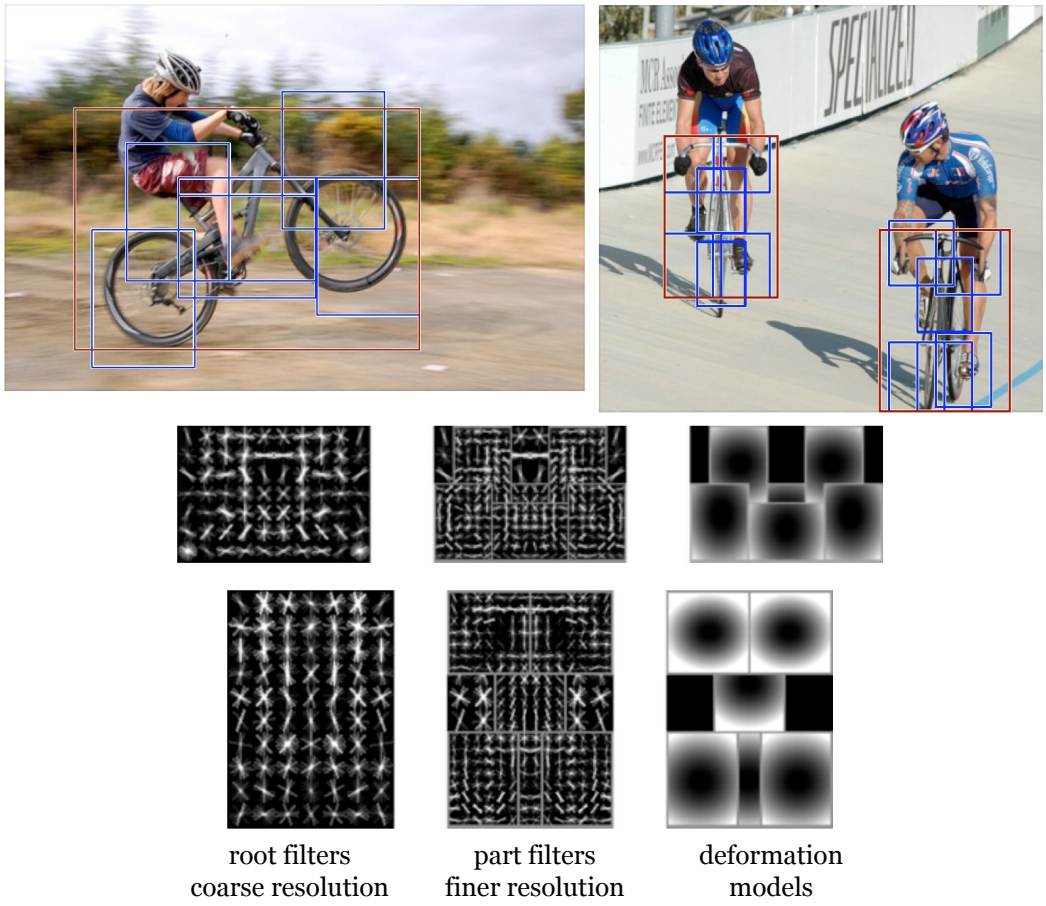

Figure 2.3: Detections obtained with a two component bicycle model of DPM [Felzenszwalb *et al.* 2010b].

locations.

[Dalal & Triggs 2005] use a sliding window approach based on a single rigid template and Histogram of Oriented Gradients (HOG) descriptors to build a detection model for pedestrian in images and videos. The pedestrian model is trained using linear SVM based on bounding boxes from positive images, and negative window set from person-free images.

2.3.1.1 Deformable part-based models

The deformable part-based models (DPM) [Felzenszwalb *et al.* 2010b] extend the single template model to mixture of deformable part-based models to handle small shape deformations, pose and viewpoint variations. The key idea behind de-

formable parts is to represent an object model using a lower-resolution *root* template, together with a set of spatially flexible high-resolution *part* templates. Each part captures local appearance properties of an object, and the deformations are characterized by links connecting them. Figure. 2.3 shows an example of detections obtained by a 2 component bicycle model of DPM. This example illustrates the importance of deformations mixture models. In this model the first component captures sideways views of bicycles while the second component captures frontal and near frontal views.

In the standard fully supervised DPMs framework, the root filter is initialized with the positive ground-truth object bounding box, and is allowed to move around in its small neighborhood to maximize the filter score. The locations of object parts are always treated as latent information due to the unavailability of object part annotations upon most occasions. A *latent* SVM (LSVM) is adopted to learn object deformation, which can alternate between fixing *latent* variables (*e.g.* , object part locations, instance-component membership) for positive examples and optimizing its objective function. The overall score of each root location p_0 is based on the best placement of all parts (*i.e.* $p_1, \ldots, p_n$):

$$score(p_0) = \max_{p_1,\ldots,p_n} score(p_0, \ldots, p_n) \tag{2.2}$$

Figure. 2.4 shows the matching process at one scale. Responses from the root and part filters are computed a different resolutions in the feature pyramid. The transformed responses are combined to yield a final score for each root location. The responses and transformed responses for the "head" and "right shoulder" parts are shown. The combined scores clearly show two good hypothesis for the object at this scale.

The deformable part model is the foundation of several champion systems for object detection challenges in PASCAL VOC 2007-2011 [Everingham *et al.* 2010]. It is successfully extended to many related tasks such as face detection [Zhu & Ramanan 2012], pedestrian detection [Xu *et al.* 2014], human pose estimation [Yang & Ramanan 2013]. DPM HSC [Ren & Ramanan 2013] replaces HOG

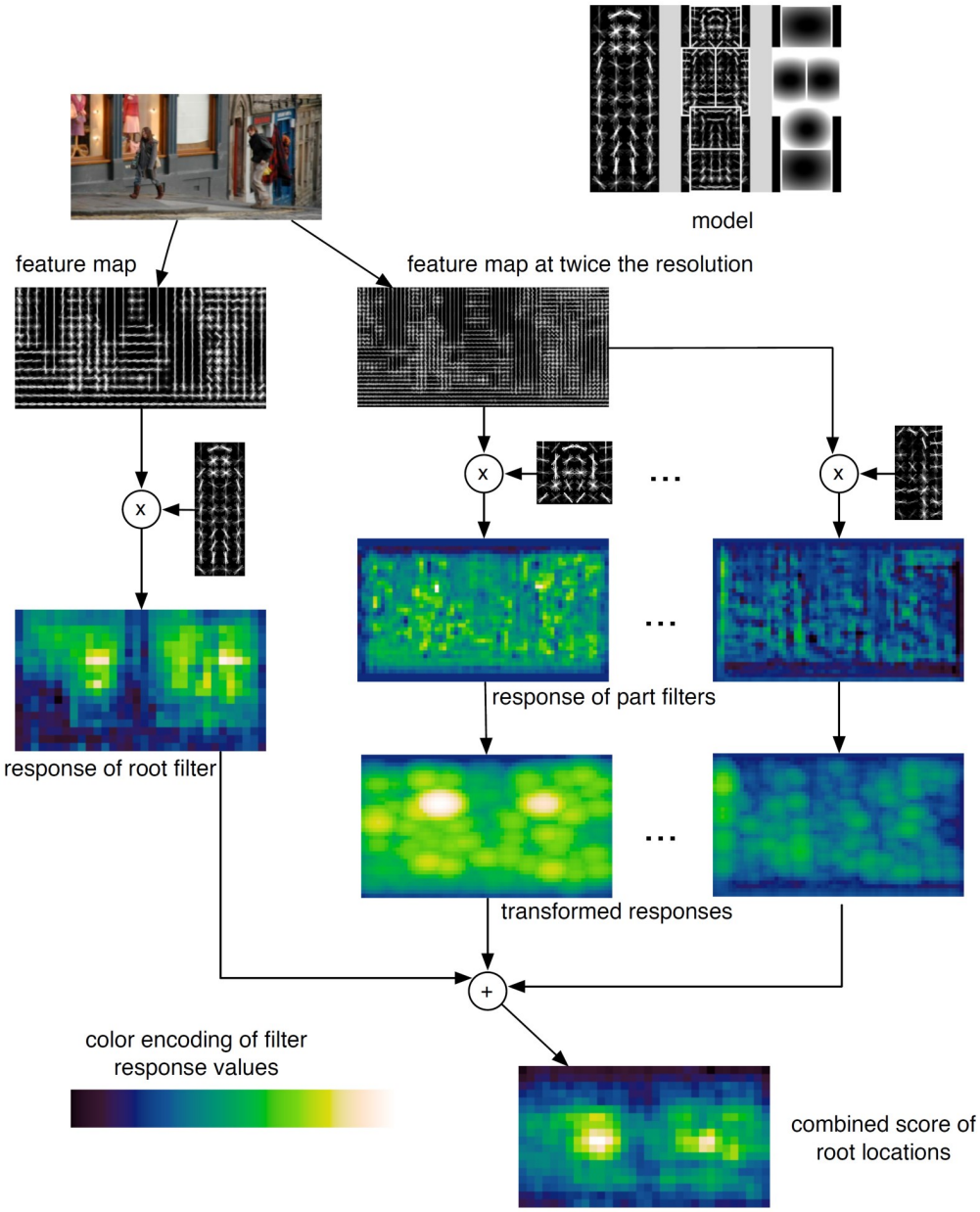

Figure 2.4: The matching process of DPM at one scale. Responses from the root and part filters are computed a different resolutions in the feature pyramid. The transformed responses are combined to yield a final score for each root location. Figure from [Felzenszwalb *et al.* 2010b].

with histograms of sparse codes (HSC), which learns sparse code dictionaries to significantly improve object detection accuracy. [Trulls *et al.* 2014] propose propose to combine bottom-up segmentation (in the form of superpixels computed

at different scales) with DPM to 'clean up' HOG features, for both root and part filters.

DPM and its various variants have advantages in handling large appearance variations for challenging datasets, however, the speed is a bottleneck of DPMs in real-time application, where speed is often considered as important as accuracy. The speed constraint mainly comes from the correlation between a sequence of root & part filters and HOG features at exhaustive sliding window locations. Some works accelerated DPM using cascade [Felzenszwalb *et al.* 2010a], coarse-to-fine [Pedersoli *et al.* 2011], branch-and-bound [Kokkinos 2011], fast Fourier transform (FFT) [Dubout & Fleuret 2012], discriminative low rank root filter with neighborhood aware cascade [Yan *et al.* 2014].

2.3.2 Region Proposal Based Approaches

Sliding window classifiers scale linearly with the number of test windows. Typically, a single-scale detection requires classifying around 10^4 – 10^5 windows per image, and the number of windows grows by an order of magnitude for multi-scale detection. Recent object detection datasets such as PASCAL VOC [Everingham *et al.* 2010], ILSVRC [Russakovsky *et al.* 2015], COCO [Lin *et al.* 2014b] adopt the IoU (Intersection over Union) to evaluate the predicted windows, which require more accurate predictions matching the aspect ratio of the objects, further increasing the search space to 10^6 to 10^7 windows per image [Hosang *et al.* 2014].

To keep the computational cost feasible, recently, class-independent region proposal (or object proposal, detection proposal) attracts a lot of attention to considerably reduce computation compared to the (dense) sliding window detection framework by generating smaller number of candidate proposals that may contain objects.

Given an image, a region proposal method aims to generate a set of candidate detection windows (around 10^3 - 10^4) that are likely to contain the objects with a high recall, under the assumption that all objects of interest share common visual

properties that distinguish them from the background. Region proposal methods are mostly based on low-level image features to generate candidate windows. Very recently, some work [Ren *et al.* 2015a, Liu *et al.* 2016] utilize deep convolutional feature maps to generate region proposals.

Two general approaches have been proposed for generating object proposals in recent years: *grouping methods* such as Selective Search [Uijlings *et al.* 2013], Constrained Parametric Min-Cuts (CPMC) [Carreira & Sminchisescu 2012], Multiscale Combinatorial Grouping (MCG) [Arbeláez *et al.* 2014] and *window scoring methods* such as Objectness [Alexe *et al.* 2012], BING [Cheng *et al.* 2014], EdgeBoxes [Zitnick & Dollar 2014]).

Among these object proposal methods, Objectness [Alexe *et al.* 2012] is one of the earliest and well known proposal methods. The objectness model is trained on a small set of training examples of mixed object categories. An initial set of proposals is selected from salient locations in an image, these proposals are then scored according to multiple cues including color contrast, edge density, location, size, and the strong "superpixel straddling" cue. Non-Maximum Suppression (NMS) is adopted to sample the initial set of candidate windows to sample high scored windows and cover diverse image locations. It is shown that more than 90% of the object instances in PASCAL VOC detection datasets can be covered (recall $>$ 90%) by these region proposals by sampling around 1000 windows per image, which is far fewer than that of sliding window approach. Therefore, a major advantage of this approach is that it enables utilization of the complex recognition models that are otherwise too slow or incompatible with the aforementioned localization strategies.

Another popular region proposal method is Selective Search proposed by [Uijlings *et al.* 2013]. It greedily merges superpixel segments with similar color and texture descriptor to generate a hierarchical segmentation tree (See Figure. 2.5). The bounding boxes of the resulting segments in the hierarchy are collected as candidate windows. This method has no learned parameters, instead features and similarity functions for merging superpixels are manually designed. Selective Search region proposal has been broadly used by many state-of-the-art object detectors,

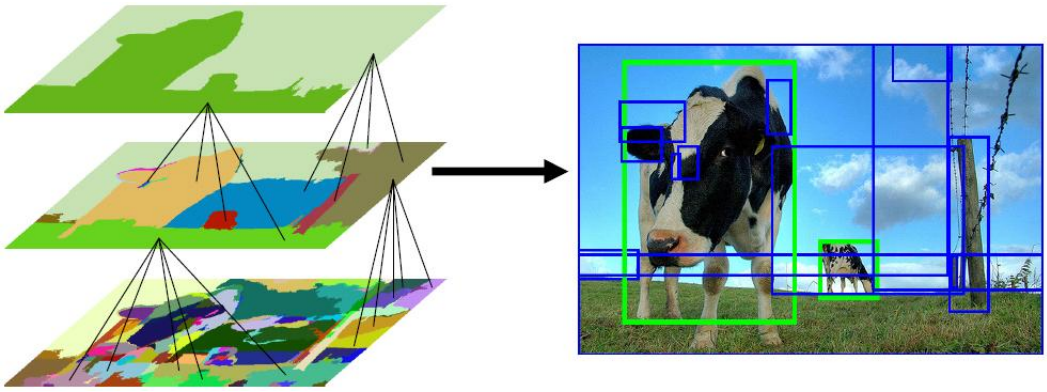

Figure 2.5: An illustration of the Selective Search region proposal method [Uijlings *et al.* 2013].

including the deep learning based R-CNN [Girshick *et al.* 2014, Girshick *et al.* 2016] and Fast R-CNN [Girshick 2015] detectors. It has a recall rate of 98% on PASCAL VOC and 92% on ImageNet with around 2000 region proposals per image.

[Hosang *et al.* 2016] discuss common strengths and weaknesses of ten recent region proposal methods, and give insights and metrics for choosing and tuning proposal methods.

2.3.2.1 Region-based convolutional neural networks

Convolutional neural networks achieved great success first in large scale image classification. Thereafter, researchers began to investigate how can the CNN be made to work as an object detector [Sermanet *et al.* 2014, Szegedy *et al.* 2013, Girshick *et al.* 2014]. Currently, the state-of-the-art object detection systems [Girshick *et al.* 2014, He *et al.* 2015, Ren *et al.* 2015a, Ren *et al.* 2015a, Liu *et al.* 2016] are deep learning (CNN) based frameworks, which benefit from the high-level features and rich representations learned by CNN.

R-CNN (Region-based Convolutional Neural Networks) [Girshick *et al.* 2014, Girshick *et al.* 2016] is one of the first deep learning based object detection frameworks which adopt region proposals. It consists of three main modules: (1) class-independent region proposals, (2) convolutional neural networks, and (3) a set of class-specific linear SVMs. In R-CNN, a CNN is first pre-trained on a large scale

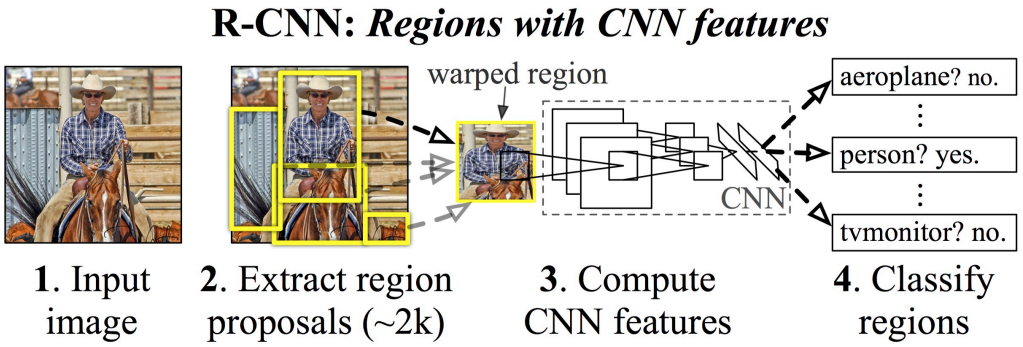

Figure 2.6: R-CNN [Girshick *et al.* 2014] object detector system overview. The system (1) takes an input image, (2) extract around 2,000 bottom-up selective search region proposals, (3) computes features for each region proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMS. R-CNNs achieves a mean average precision (mAP) of 53.7% on PASCAL VOC 2010. For comparison, [Uijlings *et al.* 2013] reports 35.1% mAP using the same region proposals, but with a spatial pyramid and bag-of-visual-words approach. The popular deformable part-based models (DPM) [Felzenszwalb *et al.* 2010b] perform at 29.6%.

dataset (*e.g.* , ImageNet 2012 classification dataset [Russakovsky *et al.* 2015] with 1,000 object categories). For each image in the target dataset, about 2,000 selective search [Uijlings *et al.* 2013] region proposals are extracted per image. The pre-trained CNN is then fine-tuned by the warped region proposals from the training images in the target dataset. This fine-tuned CNN can act as a feature extractor (*e.g.* , pool5, fc6 or fc7 output as feature vector) to extract features from warped regions with fixed length feature vectors (*e.g.* , 4096-dimensional features for fc6 and fc7 output). Finally, class-specific linear SVMs can be trained and used as classifiers to classify each region proposal from test images into target or non-target class. Non-maximum Suppression (NMS) sampling is used to discard near duplicate detected windows.

SPP-net [He *et al.* 2015] introduce a spatial pyramid pooling (SPP) layer into R-CNN to eliminate the requirement of a fixed-size input image (*e.g.* , 224×224 warped region in R-CNN). The requirement of fixed sizes is only due to the fully connected layers that demand fixed-length vectors as inputs, while the convolutional layers accept inputs of arbitrary sizes. SPP partitions the convolutional fea-

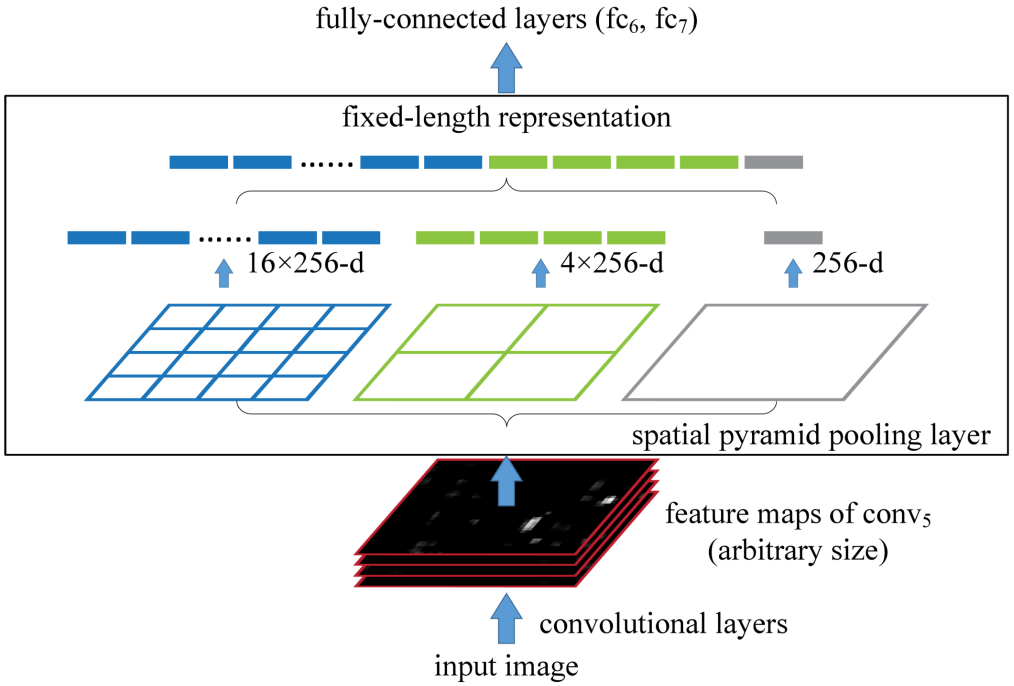

Figure 2.7: SPP-net [He *et al.* 2015] network structure with a spatial pyramid pooling (SPP) layer. Here conv5 is the last convolutional layer, and 256 is the filter number of the conv5 layer.

ture maps into divisions from finer to coarser levels, and aggregates local features in them to generate a fixed-length output regardless of the input size. Therefore, feature maps can be computed from the entire image only once by using SPP-net, rather than forwarding about 2,000 overlapping image regions for computation for each image. In practice, multiple SPP layers can exist in a network. Features for each (selective search) region proposals are extracted from the conv5 feature map of the full image. The SPP-net-based system built upon the R-CNN pipeline computes features 24 to 102 times faster than R-CNN, while has better or comparable accuracy.

A main drawback of SPP-net is that the parameters below the SPP layer can not be updated during backpropagation in training, only classifier layers (fully connected layers) are fine-tuned. However, training the convolutional layers is important for very deep networks (it was not that important for the smaller *AlexNet* and *ZF-Net*). Moreover, similar to R-CNN, SPP-net detection is also a multi-stage

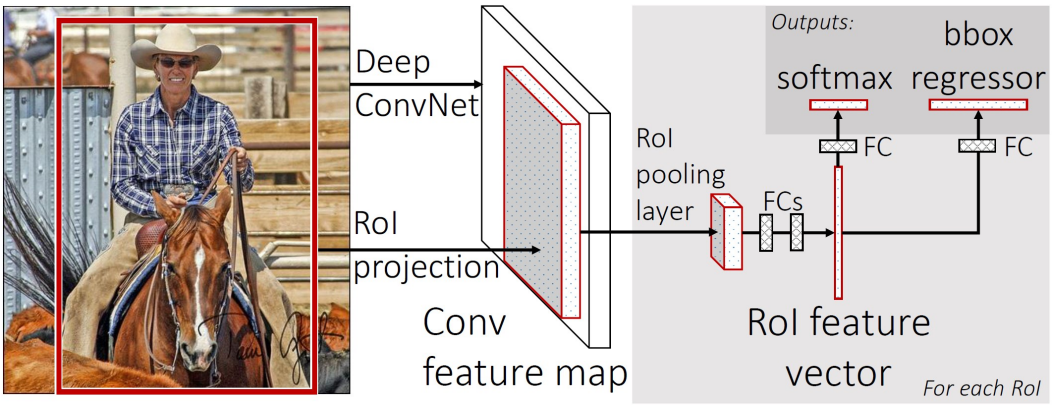

Figure 2.8: Fast R-CNN [Girshick 2015] object detector system overview. An input image and multiple regions of interest (RoIs) are input into a convolutional network. Each RoI is pooled into a fixed-size feature map and then mapped to a feature vector by fully connected layers. The network has two output vectors per RoI: softmax probabilities and per-class bounding-box regression offsets. The architecture is trained end-to-end with a multi-task loss.

pipeline that involves extracting features, fine-tuning a network with log loss, training SVMs, and finally fitting bounding-box regressors. Features are written to disk thus relatively slow. Fast R-CNN [Girshick 2015] solves these problems by proposing one network with two loss branches: (1) softmax classifier and (2) linear bounding-box regressors. The overall loss is the sum of the two loss branches. It takes in an entire image, and then passes it to the convolutional network to create a feature map. For each region proposal, it finds the corresponding local feature map. On top of that a single layer of SPP is applied which is called the RoI (region of interest) pooling layer (as opposed to multiple layers SPP that is applied in SPP-net). Then multitask loss is calculated based on bohtn the softmax classifier and bounding box regressors. This single-stage training mechanism using a multi-task loss makes the convolutional layers also trainable. Moreover, no disk storage is required for feature cashing. Training Fast R-CNN with the very deep *VGG-16* network is 9 times faster than R-CNN, while at test time is 213 times faster. Fast R-CNN achieves a higher mAP on PASCAL VOC 2012. Compared to SPP-net, Fast R-CNN trains *VGG-16* 3 times faster, tests 10 times faster, and is more accurate.

Fast R-CNN achieves near real-time rates using very deep ConvNets, when ignoring the time spent on generating region proposals (*e.g.* , Selective Search). Re-

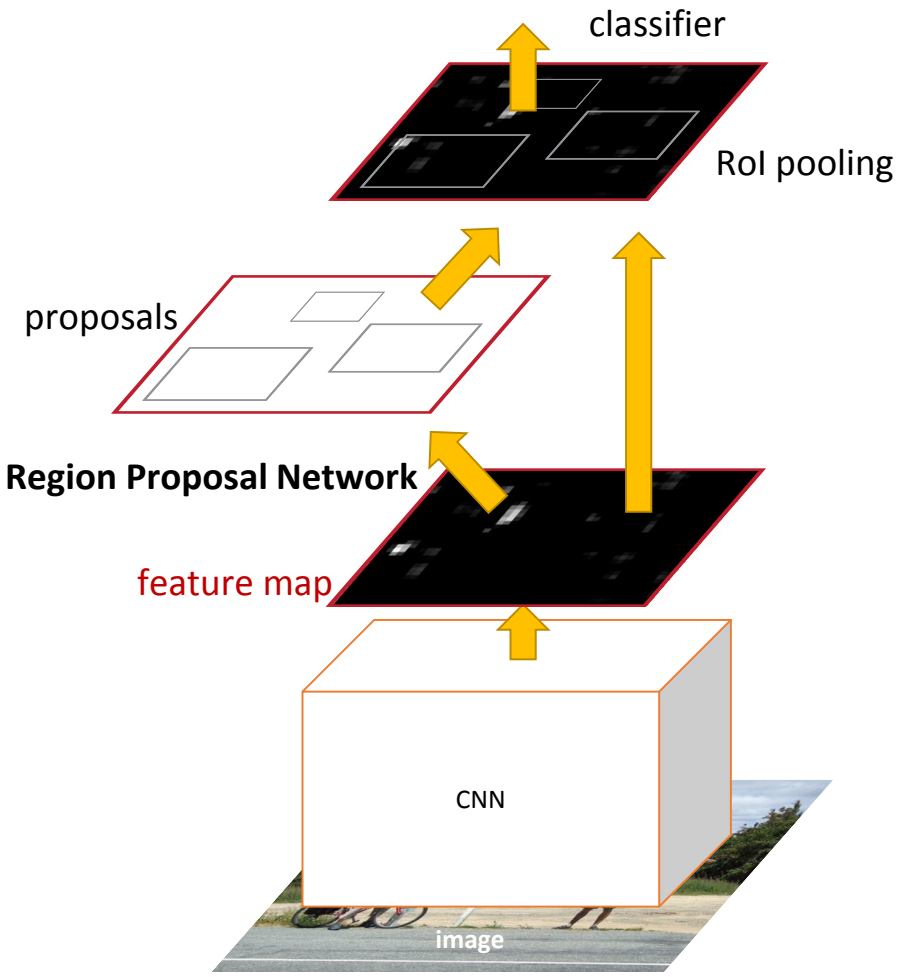

Figure 2.9: Faster R-CNN [Girshick 2015] object detector system overview. Faster R-CNN is a single, unified network for object detection. The RPN (Region Proposal Network) module serves as the 'attention' of this unified network.

gion proposals are the test-time computational bottleneck in many detection systems. In Faster R-CNN [Ren *et al.* 2015a] (Figure. 2.9), Region Proposal Networks (RPNs) are introduced for efficient and accurate region proposal generation. By sharing convolutional features with the down-stream detection network (Fast R-CNN), the region proposal step is nearly cost-free (Figure. 2.10). A Region Proposal Network is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. The RPN is trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. Faster R-CNN enables a unified, deep-learning-based object detection system

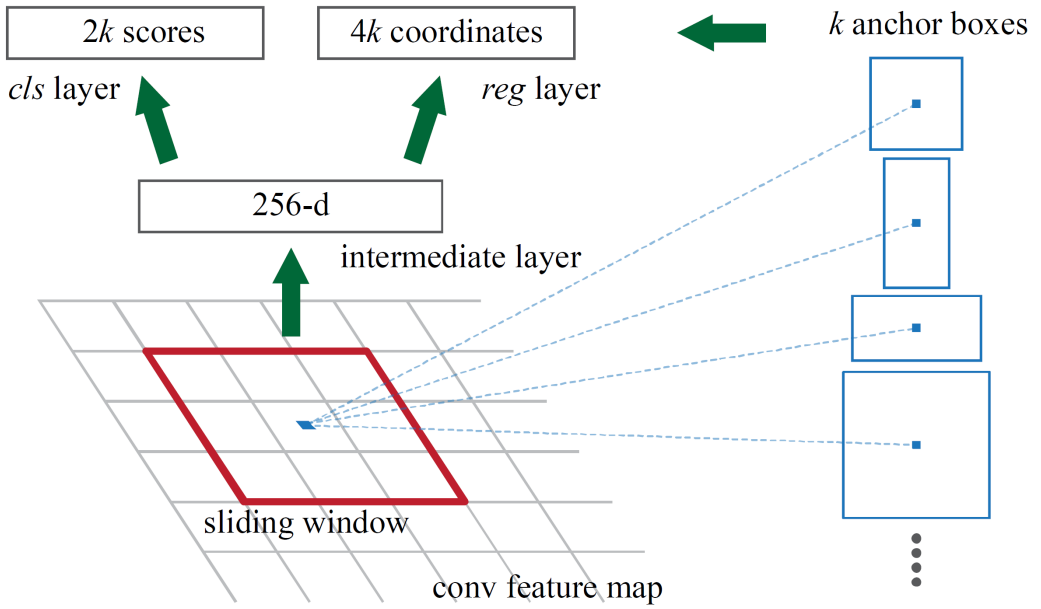

Figure 2.10: An illustration of the Region Proposal Network (RPN) [Ren *et al.* 2015a] to generate region proposals.

to run at near real-time frame rates.

All the aforementioned object detectors take a fully supervised approach, in which all the training images are annotated with bouding boxes indicating the category label and location of interesting objects.

2.4 Weakly Supervised Object Detection

Although localized object annotations are extremely valuable for object detection systems, the process of manually annotating object bounding boxes is extremely laborious, time consuming and unreliable (subjective to human bias), especially for very large-scale datasets. However, it is usually much easier to obtain annotations at *image* level (*e.g.* , from user-generated tags on Flickr or Web queries). Weakly supervised object detection aims to learn recognition models relying on training images with incomplete ground-truth bounding box annotations, given only image level (binary) labels indicating the presence or absence of object instances in the images.

Weakly supervised object detection has attracted increasing attention in recent

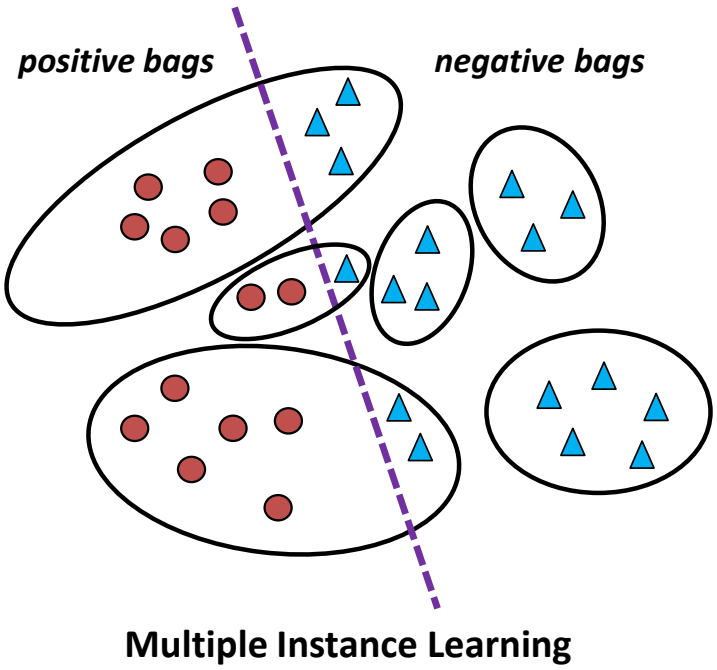

Figure 2.11: An illustration of Multiple Instance Learning (MIL) problem. In the MIL framework, there are some positive and some negative bags. A bag is positive when it has at least one positive instance, while it is negative if all the instances are negative. The objective of MIL is to train a classifier which can correctly classify a test image window as either positive or negative.

years. Based on weakly annotated examples, the common practice is to jointly learn an appearance model together with the latent object location. The majority of related work considers weakly supervised object detection as a multiple instance learning (MIL) [Maron & Ratan 1998] problem. In the MIL framework, there are some positive and some negative bags (see Figure. 2.11). A bag is positive when it has at least one positive instance, while it is negative if all the instances are negative. The objective of MIL is to train a detector (or classifier) which can correctly classify a test image window as either positive or negative. MIL problems are usually solved by finding a local minimum of non-convex objective functions (*e.g.*, MI-SVM [Andrews *et al.* 2003]). [Galleguillos *et al.* 2008] first use the MIL model to recognize and localize objects based on multiple stable segmentations. [Nguyen *et al.* 2009] and [Siva & Xiang 2011] use variants of MIL to learn object detectors from weakly labeled images and videos. Typically, the number of examples

per bag is manageable for MIL methods when utilizing region proposal methods such as Objectness [Alexe *et al.* 2012], Selective Search [Uijlings *et al.* 2013], Edge-Boxes [Zitnick & Dollar 2014]. However, it remains challenging for an algorithm to select detection windows across a large number of images. An iterative weakly supervised learning method could typically find only a local optimum depending on the initial windows. Therefore, both initialization and iterative weakly supervised learning methods are significant for the detection performance.

An other line of strategy is to exploit knowledge transfer from various domains to help weakly supervised learning for detection. Transfer learning (TL) [Shao *et al.* 2015] aims to transfer knowledge across different domains or tasks.

In this section, we will review the weakly supervised learning methods for object detection.

2.4.1 Initialization Strategies

A good number of different initialization strategies for training MIL detectors have been proposed in the literature. A simple strategy is random initialization [Kim & Torralba 2009, Pandey & Lazebnik 2011], which is to initialize randomly from relatively large windows in positive images that cover most content of the full images. [Pandey & Lazebnik 2011] modify the fully supervised Deformable Part-based Models in a weakly supervised manner without object level annotations for scene recognition and object detection: this treats the location of root filter and part filters fully latent and learns structural object detectors based on the entire image. Root filter location is initialized randomly, based on a window that has at least 40% overlap with the positive training image, while its aspect ratio is initialized roughly to the average of the aspect ratios of positive training examples. By random initialization, the object detector tends to learn spurious models of other classes or background regions, leading to lower accuracy during testing.

Another strategy is to leverage the category-independent visual saliency to avoid exhaustive search from the images. A salient region in an image should be more likely to contain object of interest than background [Borji *et al.* 2015].

Some methods use a category-independent measure which aims to predict whether an image window contains an object of interest or not. For instance, [Deselaers *et al.* 2012] generate candidate windows based on the Objectness region proposal method [Alexe *et al.* 2012] and assign per-window weights using a saliency model trained on a small meta-training set of non-target object categories. [Cinbis *et al.* 2014] rely on the category-independent Selective Search windows [Uijlings *et al.* 2013], and propose a multi-fold training procedure for MIL. To get rid of bad local minima, [Song *et al.* 2014a] initialize the object locations via a discriminative submodular covering method. [Wang *et al.* 2015] propose to cluster the Selective Search windows into sub-categories using the probabilistic Latent Semantic Analysis (pLSA), and then learn the latent categories by selecting the most discriminative subcategory for each object category. [Bilen *et al.* 2015] formulate to jointly learn a discriminative model and enforce the similarity of the selected object regions via a discriminative convex clustering algorithm.

Some methods adopt the category-specific initialization strategies. For example, [Siva & Xiang 2011] propose to initially select a single window from Objectness region proposals [Alexe *et al.* 2012] per image so that the selected windows maximize the objective function which based on intra-class and inter-class pairwise similarities. This is based on the fact that an image region containing an object instance should be similar with the regions containing the same category of objects in other images, while an image region should be dissimilar with any regions that are from negative images that are known to not contain the object of interest. [Siva *et al.* 2012] later propose a simplified method to maximize the distance between a selected candidate windows and its nearest neighbor among windows from negative images.

2.4.2 Iterative Learning Strategies

To improve the initial localization in the training image, an iterative learning approach is typically employed to generate more accurate detector. For example, [Deselaers *et al.* 2012] employ a Conditional Random Field (CRF)

[Lafferty *et al.* 2001] based model that jointly infers object hypotheses across all positive training images, by exploiting a fully-connected graphical model that encourages visual similarity across all selected object hypotheses. Except for the pairwise function, the CRF-based model is accompanied also by a unary potential function that scores candidate windows individually. The parameters of the pairwise and unary potential functions are updated and the positive windows are selected in an iterative manner. [Pandey & Lazebnik 2011] propose to iteratively run DPM detectors based on the prediction of last iteration to obtain better detection performance. However, running detectors iteratively is time-consuming for expensive detectors. [Cinbis *et al.* 2014] divide positive training images randomly into multiple fold and perform MIL iteratively on different combinations of multiple folds to avoid quickly converging to poor local optima. *et al.* [Oquab *et al.* 2015] develop a weakly supervised CNN end-to-end learning pipeline that learns from complex cluttered scenes containing multiple objects by explicitly searching over possible object locations and scales in the image, which can predict image labels and coarse locations (but not exact bounding boxes) of objects. [Bilen & Vedaldi 2016] propose a Weakly Supervised Deep Detection Network (WSDNN) method that extends a pre-trained network to a two-stream CNN: recognition and detection. The recognition and detection scores for region proposals are aggregated to predict the object category.

2.4.3 Transfer Learning Strategies

MIL-based methods tend to get stuck in local optima. Hence, a number of researchers propose to transform the easily obtained image classifiers into object detectors by transferring knowledge from external categories or other domains.

Transfer learning (TL) [Shao *et al.* 2015] aims to transfer knowledge across different domains or tasks. Two general categories of TL have been proposed in previous work: *homogeneous* TL [Donahue *et al.* 2013, Oquab *et al.* 2014, Hoffman *et al.* 2014] in a single domain but with different data distributions in training and testing sets, and *heterogeneous* TL

[Rochan & Wang 2015, Shu *et al.* 2015, Zhu *et al.* 2011] across different domains or modalities. [Shi *et al.* 2012] formulate a ranking based transfer learning method, which effectively transfers a model for predicting object location from an auxiliary dataset to a target dataset with completely unrelated object categories, which is a homogeneous TL problem. [Hoffman *et al.* 2014] propose LSDA (Large Scale Detection through Adaptation), which treats the transfer from classifiers to detectors as a homogeneous TL problem as the data distributions for image classification (whole image features) and object detection (image region features) are different. The adaptation from a classifier to a detector is however restricted to the visual domain. LSDA learns the difference between the CNN parameters of the image classifier and object detector of a "fully labeled" category, and transfers this knowledge to CNN classifiers for categories without bounding box annotated data, turning them into detectors. For LSDA, auxiliary object-level annotations for a subset of the categories are required for training "strong" detectors. [Rochan & Wang 2015] propose an appearance transfer method by transferring semantic knowledge (heterogeneous TL) from familiar objects to help localize novel objects in images and videos. [Singh *et al.* 2016] transfer tracked object boxes from weakly labeled videos to weakly labeled images to automatically generate pseudo ground-truth bounding boxes. Our work integrates knowledge transfer via both visual similarity (homogeneous TL) and semantic relatedness (heterogeneous TL) to help convert classifiers into detectors. [Shu *et al.* 2015] propose a weakly-shared Deep Transfer Network (DTN) that hierarchically learns to transfer semantic knowledge from web texts to images for image classification, building upon Stacked Auto-Encoders [Bengio *et al.* 2007]. DTN takes auxiliary text annotations (user tags and comments) and image pairs as input, while our semantic transfer method only needs image-level labels.

Recently, there exist some works [Tang *et al.* 2014a, Cho *et al.* 2015, Li *et al.* 2016] focus on the problem of unsupervised object detection through co-localization, which further alleviates the need for annotations, requiring only a set of images each containing some common object to be localized. In object co-localization, we

do not know which objects are contained in the image set, and no negative images or images known not to contain the object are provided. Co-localization outputs bounding boxes as weakly supervised localizations without strong supervision. [Tang *et al.* 2014a] proposes a joint optimization of the prior, similarity, and discriminability of both images and boxes. The proposed formulation is capable of accounting for noisy annotations in real-word images.

Chapter 3

Weakly Supervised Learning of Deformable Part-Based Models for Object Detection via Region Proposals

---

Contents

3.1 Introduction **45**

3.2 Fusing Generic Objectness and Deformable Part-Based Models for Weakly Supervised Object Detection **48**

3.2.1 Object Estimations: Initialization 49

3.2.1.1 Region extraction 50

3.2.1.2 Salient reference region 50

3.2.1.3 Coarse candidate window pool 51

3.2.1.4 Object invariant estimations 51

3.2.2 Learning Latent Object Classes via Region Classification . . . 54

3.2.2.1 Region representation 55

3.2.2.2 Region classification 55

3.2.3 Weakly Supervised DPMs Training and Testing Details 57

3.2.3.1 Single region initialization for weak DPMs (S-WDPMs) detection 57

3.2.3.2 Multiple region initialization for weak DPMs (M-WDPMs) detection 57

3.2.4 Bounding Box Post-processing 59
3.3 Experimental Evaluation 61
3.3.1 Experiments with S-WDPMs on PASCAL VOC Subsets 62
3.3.1.1 Datasets and settings 62
3.3.1.2 Evaluation protocol 63
3.3.1.3 Experimental evaluation 63
3.3.2 Experiments with M-WDPMs on PASCAL VOC 66
3.3.2.1 Dataset and settings 66
3.3.2.2 Parameter selection 68
3.3.2.3 Annotation evaluation 69
3.3.2.4 Detection evaluation 72
3.3.2.5 Error analysis 74
3.3.2.6 Running time 76
3.3.3 Preliminary Results with M-WDPMs on MS COCO 78
3.4 Summary 79

The success of deformable part-based models (DPMs) for visual object detection relies on a large number of labeled bounding boxes. With only image-level annotations, our goal is to propose a model enhancing the weakly supervised DPMs by emphasizing the importance of location and size of the initial class-specific root filter. To adaptively select a discriminative set of candidate bounding boxes as this root filter estimate, first, we explore the generic objectness measurement to combine the most salient regions and "good" region proposals. Second, we propose learning of the latent class label of each candidate window as a binary classification problem, by training category-specific classifiers used to coarsely classify a candidate window into either a target object or a non-target class. Moreover, we incorporate the contextual information from image classification, by combining the image-level classification score with object-level DPM detection score, to obtain a final score for detection. Finally, we design a flexible enlarging-and-shrinking post-processing procedure to modify the DPMs outputs, which can effectively match the

approximative object aspect ratios and further improve final accuracy. Extensive experimental results on the challenging PASCAL Visual Object Class (VOC) 2007 and the Microsoft Common Objects in Context (MS COCO) 2014 dataset demonstrate that our proposed framework is effective for initialization of the DPMs root filter. It also shows competitive final localization performance with state-of-the-art weakly supervised object detection methods, particularly for the object categories which are relatively salient in the images and deformable in structures.

3.1 Introduction

Object detection/localization in images/videos is one of the most widely studied problems in computer vision applications [Zhang *et al.* 2010, Zhu *et al.* 2014, Girshick *et al.* 2014] with the explosive growth of online images/videos today. It can also be extended to numerous applications related to the multimedia community, *e.g.*, image and video retrieval, video surveillance [Foresti *et al.* 2002, Nascimento & Marques 2006], traffic safety: self or assisted driving systems, *etc*. This task remains challenging mainly due to scale and viewpoint variation, deformation, occlusion, background clutter, intra-class variations and inter-class similarities for objects in real world images/videos. For most of the existing methods, a fully supervised learning (FSL) approach is adopted [Dalal & Triggs 2005, Felzenszwalb *et al.* 2010b, Szegedy *et al.* 2013, Zhu *et al.* 2014, Girshick *et al.* 2014], where positive training images are manually annotated with bounding boxes encompassing the objects of interest. This manual annotation of object location for large-scale image databases is extremely laborious and unreliable though quite valuable for learning accurate object detectors. However, it is usually far easier to obtain weakly labeled data, where image-level labels (*e.g.*, user generated image tags on Internet) are presented. For example, the recently popular ImageNet ILSVRC dataset [Russakovsky *et al.* 2015] contains far fewer object-level annotations (bounding boxes) than image-level labels. As a result, in our work, in contrast to the traditional FSL, we are concerned with weakly supervised learning (WSL) for object detection, where the exact object locations in positive training examples are

not provided, giving only the binary labels indicating the presence or absence of the objects of interest.

Deformable part-based models (DPMs) [Felzenszwalb *et al.* 2010b] and their variants [Girshick *et al.* 2011, Azizpour & Laptev 2012, Ren & Ramanan 2013] have achieved remarkable success in supervised object detection on challenging PASCAL VOC datasets [Everingham *et al.* 2010] for a long period. The DPMs represents an object with a holistic *root* filter that approximately covers an entire object and with several higher resolution *part* filters that capture smaller local appearances (parts) of the object. It also characterizes the deformations by links connecting different parts. In the standard (fully supervised) DPMs framework, the root filter is initialized with the positive ground-truth object bounding box, and is allowed to move around in its small neighborhood to maximize the filter score. The locations of object parts are always treated as latent information due to the unavailability of object part annotations upon most occasions. A *latent* SVM (LSVM) is adopted to learn object deformation, which can alternate between fixing latent values (part locations) for positive examples and optimizing its objective function.

[Pandey & Lazebnik 2011] modify the fully supervised DPMs in a weakly supervised manner without object level annotations: this treats the location of root filter and part filters fully latent and learns structural object detectors based on the entire image. Root filter location is initialized randomly, based on a window that has at least 40% overlap with the positive training image, while its aspect ratio is initialized roughly to the average of the aspect ratios of positive training examples. However, the specific size and location of the initial root filter, as well as their aspect ratio, are indicated to have a significant impact on the final localization result [Dalal & Triggs 2005, Felzenszwalb *et al.* 2010b, Pandey & Lazebnik 2011]. By random initialization, the object detector tends to learn spurious models of other classes or background regions, leading to lower accuracy during testing. To the best of our knowledge, methods for initializing the root filter based on theoretical deduction in weakly supervised DPMs, as well as the definition of the object aspect ratios, have not been properly studied in [Pandey & Lazebnik 2011].

To make up the performance gap between weakly and fully supervised DPMs,

in this paper, our goal in this paper is to propose a model enhancing the weakly supervised DPMs by emphasizing the importance of location and size of the initial class specific root filter. To be more precise, our goal is to discover a reliable initial set of image windows that are likely to contain the target objects in the positive training images with only category level annotations, so as to represent the object instances. Hence, our WSL framework incorporates adaptive window selection from class independent object proposals and training of deformable part-based models. In particular, we explore the "objectness" approaches [Alexe *et al.* 2012, Uijlings *et al.* 2013], which generate class independent object proposals with corresponding scores indicating their probabilities of being object instances. We then adaptively select a reliable set of windows from the derived object proposals for each image as initialization, by incorporating visual saliency and "objectness" scores. Two different initialization schemes are developed: *single* region and *multiple* region initilization. The former tends to select one relative larger bounding box which may contain the most salient part in the image, while the latter is far more general, which selecting a small number of object estimations that can also capture smaller and scattered objects. For multiple region initialization, the region labels are latent information. We learn the latent class label by framing it as a classification problem, which tries to coarsely classify each region into a target object class or a non-target class by some class specific classifiers. The generated object estimations are treated as the initial root filter estimates for training DPMs detectors.

The main contributions in this work are several-fold:

1. We propose a selection model based on generic "objectness" (region proposals) and visual saliency to adaptively select a discriminative set of candidate windows which tend to represent the object instances in each weakly labeled training image.

2. We frame the learning of the latent class label of each candidate window as a binary classification problem, by training category specific classifiers, which try to coarsely classify a candidate window into either a target object or a

non-target class.

3. We incorporate the contextual information from image classification, by combining the image-level classification score with object-level DPM detection score, to improve object detection.

4. We propose to use a flexible enlarging-and-shrinking post-processing procedure to modify the predicted output of the DPMs detector, which can effectively generate more accurate bounding boxes by better conserving foreground and cropping out plain background regions, to approximatively match the object aspect ratios.

5. Extensive experiments are carried out on two subsets and on the entire set of the challenging PASCAL VOC 2007 database [Everingham *et al.* 2010] with different criteria, namely annotation accuracy in terms of correct localization on training set, and detection accuracy in terms of average precision on test set. Experimental results demonstrate that our proposed framework is effective for initialization of the DPMs root filter and that it shows shows competitive final localization performance with the state-of-the-art weakly supervised object detection methods. To the best of our knowledge, we are the first to present weakly supervised results on the Microsoft COCO 2014 dataset [Lin *et al.* 2014b].

The rest of this chapter is organized as follows: we present our weakly supervised DPMs framework in detail in Section 3.2, while in Section 3.3 we present our experimental results and the comparison with other methods on PASCAL VOC 2007 and Microsoft COCO 2014 datasets. Section 3.4 concludes this chapter.

3.2 Fusing Generic Objectness and Deformable Part-Based Models for Weakly Supervised Object Detection

In this section, we detail our approach of the weakly supervised DPMs for object detection. First, we introduce our approach to adaptively select the representative and

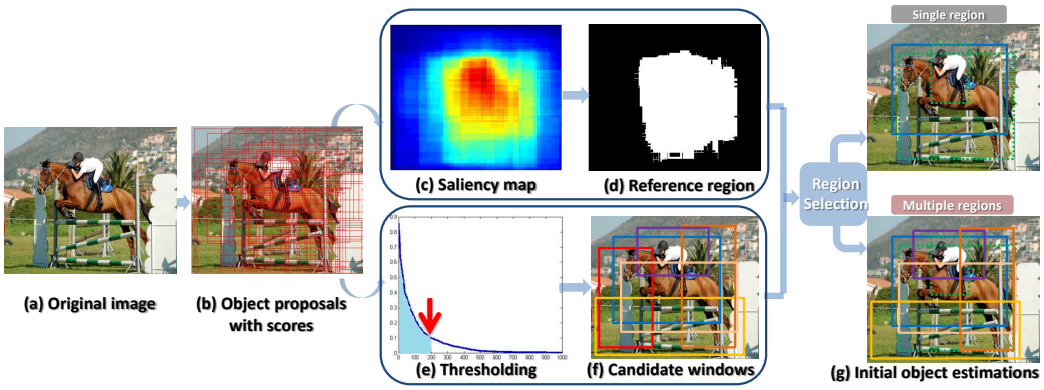

Figure 3.1: Illustration of our proposed method to extract the initial object estimations: for an input image (a), object proposals (b) and corresponding scores indicating the probabilities of containing objects are generated using the Objectness [Alexe *et al.* 2012] or Selective Search [Uijlings *et al.* 2013] method. (c) is the saliency map derived from (b), and (d) is the reference region obtained by thresholding (c). A coarse set of candidate windows (f) is selected based on the sorted scores of object proposals (e) after non-maximum suppression (NMS). In the top image of (g), which indicates the single region selection scheme, the blue window is our initial object estimation obtained by optimizing the overlap between (d) and (f). The bottom image of (g) indicates the multiple region selection scheme. Its color windows with solid lines are multiple finer regions which are assumed to represent the objects in the original image. For both images of (g), the green dot line windows are ground-truth bounding boxes for person and horse, respectively.

discriminative candidate regions from the category-independent object proposals. Second, we elaborate how to learn latent class information when multiple regions are selected. We then briefly describe the weakly supervised learning procedures using the selected regions with DPMs and the detection rescoring algorithm using classification scores as contextual information for testing. Finally, we propose our new post-processing method to further refine the predicted object bounding box obtained by a weak DPMs detector, so as to cover the object more precisely.

3.2.1 Object Estimations: Initialization

In the weakly supervised DPMs training procedure, good initialization of the root filter is crucial. Our goal is thus to discover a reliable initial set of image windows likely to contain the target objects in the positive training images with only image-level annotations, so as to represent the object instances.

3.2.1.1 Region extraction

Two general approaches have been proposed for generating class-independent object proposals in recent years: *window scoring methods* such as Objectness [Alexe *et al.* 2012], BING [Cheng *et al.* 2014], EdgeBoxes [Zitnick & Dollar 2014] and *grouping methods* such as Selective Search [Uijlings *et al.* 2013], Constrained Parametric Min-Cuts (CPMC) [Carreira & Sminchisescu 2012], Multiscale Combinatorial Grouping (MCG) [Arbeláez *et al.* 2014]). We use Selective Search since it has been used as the proposal generating method by the state-of-the-art supervised R-CNN detector [Girshick *et al.* 2014]. We also report results using the Objectness method [Alexe *et al.* 2012] to compare with prior detection work [Alexe *et al.* 2012], [Tang *et al.* 2014b].

Given an input image I (shown in Figure 3.1(a)), we first select top n scored windows $W = \{w_1, w_2, \ldots, w_n\}$ and corresponding scores, denoted as $S = \{s_1, s_2, \ldots, s_n\}$, indicating the probabilities of covering objects within them, generated by Selective Search (shown in Figure3.1 (b)). To balance a high recall (*i.e.*, covering more objects) and computation efficiency (*i.e.*, small number of region proposals), we set $n = min(1000, N)$ according to [Hosang *et al.* 2014], where N is the number of proposals generated by Selective Search.

Based on the fact that the region proposal method is designed to capture all possible objects within an image, we assume that it is sufficiently reliable to provide a set of good candidate windows $W^* \subseteq W$ covering the objects of interest. However, windows with higher scores are not always the effective choices [Shi *et al.* 2012]: they usually encompass other noisy background, or they may cover only some object parts. To extract a reliable set of object estimations from the pool of n windows, we design a sequential selection scheme shown in Figure3.1 (c)-(g).

3.2.1.2 Salient reference region

For weakly supervised learning of DPMs detectors, it is obvious that the initialization of the root filter is significant. The detector will be seriously damaged if it shoots on the background region. Consequently, it is an absolute necessity to start

from visually meaningful regions (foreground objects). Identifying visually salient regions is essentially useful in object detection. Inspired by the success of visual saliency applied in salient object recognition [Zhang *et al.* 2010, Li *et al.* 2013], we compute the reference region R (shown in Figure 3.1 (d)) by taking the threshold and merging the discrete saliency map (or heat map) M into one or more connected region(s) using [Otsu 1979] (shown in Figure 3.1 (c)). The value of saliency map M at pixel $I(i,j)$ is obtained by summing up the scores of the windows that cover this pixel:

$$M(i,j) = \sum_{k=1}^{n} M_k(i,j) \tag{3.1}$$

where,

$$M_k(i,j) = \begin{cases} s_k, & \text{if } I(i,j) \in w_k, \forall w_k \in W, \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

The reference region R can be one connected (continuous) region or several discrete regions in the image according to the score range and threshold value.

3.2.1.3 Coarse candidate window pool

It is known that the score predicted by Selective Search (*i.e.*, objectness score) corresponds to the probability of containing a target object to some extent. To take advantage of this auxiliary information, we concurrently select the top 200 scored windows out of n windows as candidates, (shown in Figure 3.1(e)). To avoid near duplicate candidate windows, we further perform non-maximum suppression (NMS) to obtain a finer set of candidates. Figure 3.1 (f) illustrates the derived smaller set of l confident candidates $\hat{W} = \{\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_l\}$ and their corresponding scores denoted as $\hat{S} = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_l\}$.

3.2.1.4 Object invariant estimations

Given the reference region R which implies the most salient region (or regions) within an image, and confident candidate windows $\hat{W}$ with scores $\hat{S}$, the overlap between them provides valuable information for finding the locations of target ob-

jects. We will propose two different schemes to fuse the salient region(s) with the extracted candidate windows.

Single region initialization: In [Pandey & Lazebnik 2011], the root filter of the DPMs is randomly initialized from a *single* window which covers at least a 40% overlap with the original image. Hence, we also filter out only one *single* window w^* from the candidate pool $\hat{\mathbb{W}}$ in order to obtain a direct comparison with [Pandey & Lazebnik 2011]. Intuitively, we expect this window estimation to have a larger overlap with the salient reference region *R*, as well as a relatively higher objectness score. Therefore, the estimation of the initial object bounding box with objectness score (w^*, s^*) (Figure 3.1(g), top image) can be determined by optimizing the following function:

$$(w^*, s^*) = \underset{\hat{w}_i \in \hat{\mathbb{W}}, \hat{s}_i \in \hat{\mathbb{S}}}{\arg\max} \; [\alpha \hat{s}_i + (1-\alpha)\frac{area(R \cap \hat{w}_i)}{area(R \cup \hat{w}_i)}], \quad i \in [1, l] \tag{3.3}$$

where α is a parameter used to control the influence of the objectness score s_i. In practice, $\alpha = 0.2$, was selected by a grid search over $\{0.1, 0.2, 0.3, 0.4\}$ on a validation set, for the purpose of emphasizing the priority of the intersection over union (IoU) overlap between the candidate window and the merged salient reference region.

The single region initialization scheme prefers to select a relatively large region which may contain the most salient part in the image. When very few objects are closely gathered in images, it can produce good DPMs object detectors in a weakly supervised manner. For example, by adopting the single region scheme, the blue window in Fig.3.1(g) top image, is used as a positive training example (*i.e.* DPMs root filter initialization) for both the *horse* and the *person* categories. Moreover, the strategy of taking large windows in positive images exploits the inclusion structure of the multiple instance learning (MIL) problem for object detection: although large windows may contain a significant amount of background features, they are likely to include positive object instances and their contextual information.

Multiple region initialization: In fact, multiple objects (*e.g.* , 2.5 objects on average for PASCAL VOC 2007 trainval dataset, 7.7 for MS COCO 2014) can be scattered anywhere in an image. We can therefore further improve DPMs detectors by providing more object estimations as root filter initialization, instead of training the object detectors with a single window for each image. For each image, we are motivated to select a small number of object estimations that can also capture smaller and scattered objects, better representing the original image.

Meanwhile, object proposal algorithms such as Selective Search and Objectness tend to generate more overlapping bounding boxes on larger objects than on smaller ones. Consequently, scattered small objects are likely to be ignored using Eq. (3.1). Hence, in order to fully consider these objects which were originally ignored by Eq. (3.1), we modified it by dividing the sum of scores by the square root of the number of windows that cover this pixel:

$$M(i,j) = \frac{1}{\sqrt{\hat{k}}} \sum_{k=1}^{n} M_k(i,j) \tag{3.4}$$

where, $M_k(i,j)$ is defined as the same in Eq. (3.2), and $\hat{k}$ is the number of windows that cover pixel $I(i,j)$. We show some heat map examples generated by Eq. (3.1) and Eq. (3.4) in Figure 3.2.

We adopt similar criteria to the score function Eq. (3.3), with the best α being set to 0.4 (for both PASCAL VOC 2007 and MS COCO 2014) from a grid search over {0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8}. Instead of only selecting the maximum scoring window in Eq. (3.3), we pick out top Q scored windows W^* for each image. We will discuss the value of Q in the experiment part.

After generating several object estimations from each image, the next step is to approximately identify the class label of each estimation given only the labels of the whole image. For example, in Fig. 3.1(g) bottom image, the color windows with solid lines are associated with the *horse* and *person* labels. However, so far we have no idea which object(s) (or even background) is/are inside each bounding box. Our goal will be to solve this problem in the next subsection.

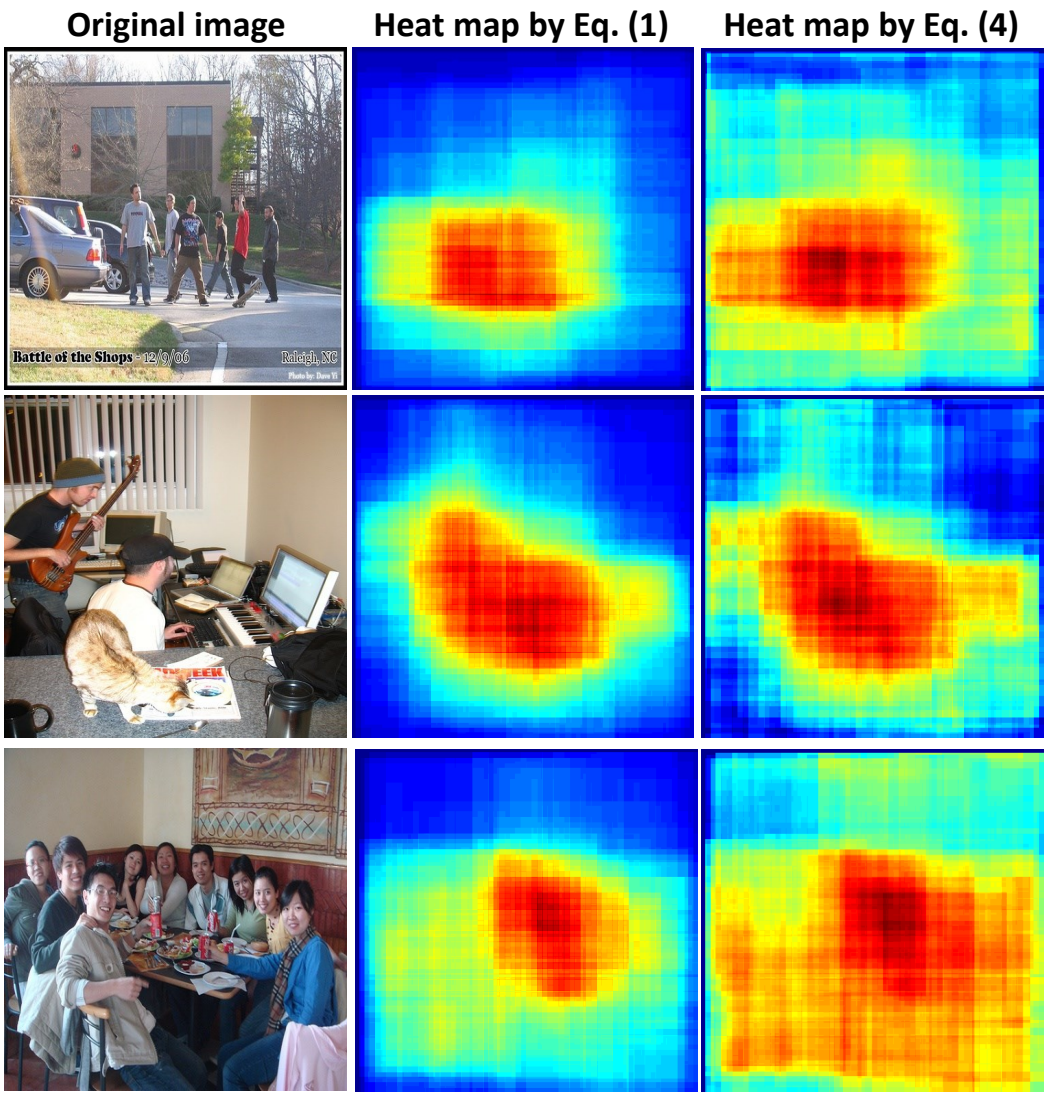

Figure 3.2: Some heat map examples generated by Eq. (3.1) and Eq. (3.4).

3.2.2 Learning Latent Object Classes via Region Classification

For each positive training image, we have generated Q object invariant estimations with the multiple region initialization scheme ($Q = 1$ for single region initialization, and we use the image-level labels as training annotations). Consider an object category (*e.g.* , *horse*), which has P positive training images, we can obtain a total number of $z = P * Q$ object estimations. Obviously, some of these object estimations come from other categories (*e.g.* , *person*, *sheep*, object parts or the background

regions as well), where the class labels are latent information. For single region initialization, the unique generated window is used to initialize the DPMs root filter for any categories appearing in the image. As for multiple region initialization, in this paper we frame the latent class learning problem as a classification problem by coarsely classifying these object estimations into either the target object category or the non-target category (*i.e.*, other classes, object parts or background).

3.2.2.1 Region representation

We use the deep convolutional neural network (CNN) features to represent the regions (object estimations). Firstly, we pre-train an eight-layer (five convolutional layers and three fully-connected layers) *Alex-Net* [Krizhevsky *et al.* 2012] CNN with *caffe* implementation [Jia *et al.* 2014] on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 classification dataset [Russakovsky *et al.* 2015], which contains 1.2 million images of 1000 categories. We then warp each region into a required fixed pixel size of 227×227, and subtract it with the mean RGB image of the training set, before forward propagating it through the network. Finally, we take the output of the $fc6$ layer as R-CNN [Girshick *et al.* 2014], which is a 4096-dimensional feature vector, to represent the input region. While this feature extraction process is similar to that of R-CNN, it is worth noticing that we do not fine-tune the pre-trained CNN on the target dataset. This is because the object level annotations are assumed not to be available in the weakly annotated data. We do not pad the region with additional image context around it either, as our region estimation is already expected to have a significant coverage of the context information due to our selection schemes in Section 3.2.1.3.

3.2.2.2 Region classification

Consider training a *horse* detector. For all the P positive training images in the *horse* category, we generate z object invariant estimations. Intuitively, only part of these z regions contains the target *horse* object, others may have *person*, *sheep*, *dog* or even background. We learn the latent categories in these regions via region

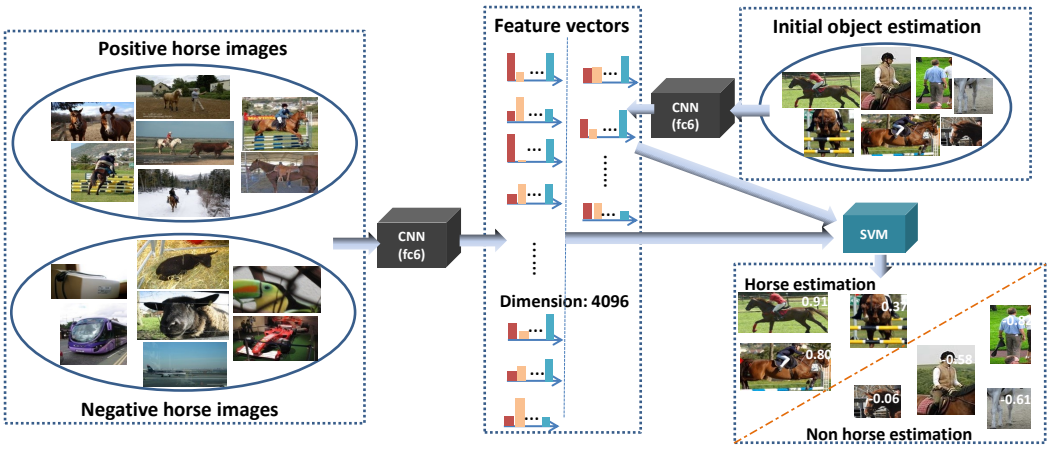

Figure 3.3: Illustration of our latent class learning framework for the *horse* category. For each object category, we train a linear SVM classifier with the CNN features (output of CNN's $fc6$ layer) of image-level samples (as shown in the left part). Object estimations from the positive training images of this category are scored by its SVM. We select the regions with higher scores by thresholding as the representative objects of this category (*horse* vs. *non horse* for this example).

classification.

We first train a *horse* linear SVM classifier [Chang & Lin 2011] using the images labeled with *horse* as positive training examples and those without *horses* as negative examples. We compute the $fc6$ 4096-dimensional CNN features as in Section 3.2.2.1 on whole images. We then run the trained *horse* classifier on the z object invariant estimations in the positive training images. By thresholding the SVM scores, finally we obtain a subset z' regions from z estimations ($z' < z$). These z' regions are assumed to represent the target *horse* category, which can be treated as positive training examples of the *horse* detector.

Suppose we have K categories that we want to detect. We train one binary SVM classifier on positive and negative images of each category, and run these K classifiers on their corresponding object estimations. We select high scoring regions for each target category so as to represent the objects of interest. Fig. 3.3 shows the latent class learning framework using SVM classification on the *horse* category.

3.2.3 Weakly Supervised DPMs Training and Testing Details

We design two different kinds of deformable part-based models for weakly supervised object detection according to different initialization schemes in Section 3.2.1.

3.2.3.1 Single region initialization for weak DPMs (S-WDPMs) detection

Similarly to [Felzenszwalb *et al.* 2010b], each root filter hypothesis in a positive training image is initialized with the corresponding derived bounding box from the single region initialization scheme. The size and aspect ratio of the DPMs root filter are decided by the average size and aspect ratio of the object estimation boxes (ground-truth bounding box and aspect ratio are used in [Felzenszwalb *et al.* 2010b]). The root filter hypothesis is allowed to move around in a small neighborhood to maximize the filter score so as to compensate for imprecise bounding box estimation from single region initialization scheme of Section 3.2.1.4. In order to obtain a direct comparison with [Pandey & Lazebnik 2011], we also represent an image by a multiscale HOG feature pyramid [Dalal & Triggs 2005] of 16 levels. For this S-WDPMs model, we use only a single component, since the multiple components are used for detecting objects with different views (S-WDPMs is trained on each view/category, *e.g.*, *Left*, *Right*). We set the number of parts in this DPMs to 8 as in [Felzenszwalb *et al.* 2010b]). For negative training examples, we use random negatives from other object categories. For testing, the sliding window approach is adopted. This single region initialized weakly supervised DPMs detection model is denoted as *S-WDPMs*. We refer the reader to [Felzenszwalb *et al.* 2010b] for more details concerning the DPMs training and detection procedures.

3.2.3.2 Multiple region initialization for weak DPMs (M-WDPMs) detection

For the M-WDPMs (multiple region initialized weakly supervised DPMs), we make it much "*deeper*" with the *DeepPyramid* feature [Girshick *et al.* 2015], for the reason that the HOG feature is suboptimal compared to deep features computed by CNN [Szegedy *et al.* 2013, Sermanet *et al.* 2014, Girshick *et al.* 2014, He *et al.* 2015,

Wang *et al.* 2015]. The feature map is computed by the fifth convolutional layer ($conv5$), which has 256 feature channels. We represent each image (or region) with a feature pyramid of 7 levels as in [Girshick *et al.* 2015]. For training, the selected object estimations from Section 3.2.2.2 are treated as positive training examples, and the random windows from negative images are defined as negative examples. We use a DPMs with 3 components and 8 parts per component according to [Girshick *et al.* 2015]. The training and testing procedures are similar to S-WDPMs above, but we add a simple bounding box rescoring stage with the help of a front-to-end CNN padded with a softmax classifier as follows.

The contextual information provided by classification and detection can mutually boost the performance of the other, based on the assumption that they adopt different information[Song *et al.* 2011, Ouyang *et al.* 2015]. Classification looks at the objects and their contextual information, while detection mainly focuses on the object shape and all parts. For example, if an object is occluded or truncated, it will be difficult for the detector while the classifier could still have enough information such as context and certain parts. Inversely, the detector is able to find small objects and objects appearing in non standard context, while the classifier may fail. Hence, we are motivated to combine the classification score and the detection score. We formulate the rescoring function as a linear combination of the DPMs detection score and region classification score:

$$s^i_{det} = \kappa s^i_{M-WDPMs} + (1-\kappa)s^i_{cls}, \quad i \in [1, K] \tag{3.5}$$

where, $0 \leqslant s^i_{M-WDPMs} \leqslant 1$ is the normalized DPMs detection score on a sub-window of the i^{th} detector, and $0 \leqslant s^i_{cls} \leqslant 1$ is the softmax classification score of the corresponding i^{th} category on this sub-window. κ is a hyper-parameter used to leverage the two scores, which ranges from 0.5 to 1.0. K is the number of object categories. The final predicted windows are obtained by thresholding the S^i_{det} in Eq. (3.5).

To train this front-to-end CNN classifier described above, we fine-tune the pre-trained CNN with image level annotations on the training set of the target dataset.

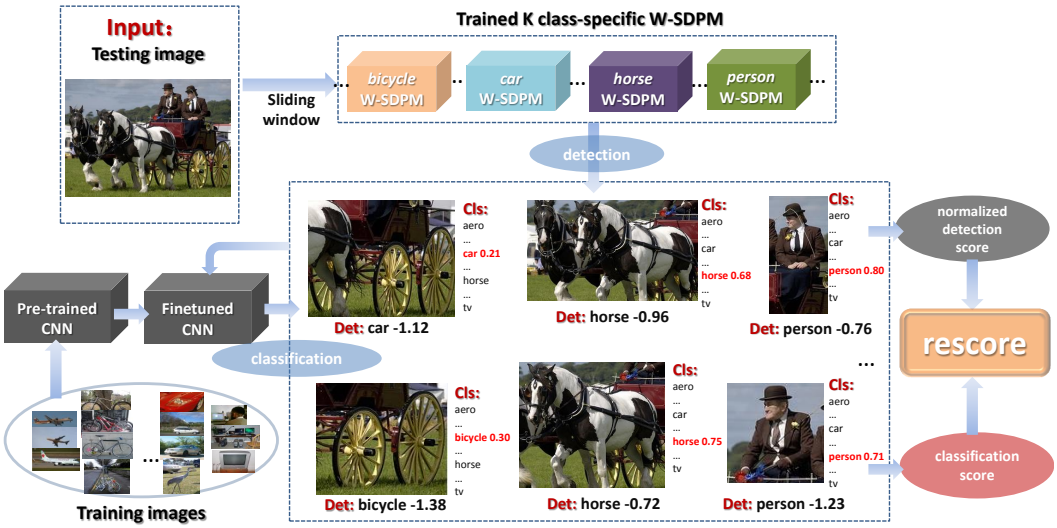

Figure 3.4: Illustration of detection rescoring using an M-WDPMs and CNN softmax classifier. For a testing image, K (number of classes in the target dataset) class-specific M-WDPMs are applied on it in a sliding window manner. For each sub-window detected by M-WDPMs, the normalized detection score is rescored by the softmax classifier of the detected category. In this example, the wrongly detected car and bicycle are finally discarded by the detector after the rescoring stage.

We implement it by removing the last 1000-way softmax layer while keeping all the other parameters and adding a new randomly initialized K-way softmax classification layer. We then fine-tune the entire network based on the image-level labels.

In [Felzenszwalb *et al.* 2010b], contextual information is exploited to rescore the bounding boxes. However, it needs object-level annotations to extract the contextual information. Our detection rescoring method does not require the object-level annotations, and leads to a remarkable improvement in average precision on several classes in the PASCAL VOC 2007 dataset (see Section 3.3.2). In [Ouyang *et al.* 2015], the image classification scores are used as contextual features, and concatenated with the object detection scores to form a final feature vector, based on which a linear SVM is learned to refine the detection score. An example of our bounding box rescoring procedure is shown in Fig. 3.4.

3.2.4 Bounding Box Post-processing

In many cases, the bounding boxes generated by DPMs detectors are too large

Algorithm 1 Bounding box post-processing pipeline.

INPUT:

Original bounding box: $w = (x_{min}, y_{min}, x_{max}, y_{max})$;
Original image width: w_o; original image height: h_o;
Maximal expanding rate: $\beta = 1.2$;
Laplacian filter shape: $\gamma = 0.2$.

OUTPUT:

Cropped bounding box: $w' = (x'_{min}, y'_{min}, x'_{max}, y'_{max})$.

centroid: $(x_c, y_c) = (\frac{x_{min}+x_{max}}{2}, \frac{y_{min}+y_{max}}{2})$
augmented width: $a = \beta * (x_{max} - x_{min})$
augmented height: $b = \beta * (y_{max} - y_{min})$
if $x_c - \frac{a}{2} > 0$ **then**
  $x^{aug}_{min} = ceil(x_c - \frac{a}{2})$
else
  $x^{aug}_{min} = 1$
end if
if $x_c + \frac{a}{2} < w_o$ **then**
  $x^{aug}_{max} = floor(x_c + \frac{a}{2})$
else
  $x^{aug}_{max} = w_o$
end if
y^{aug}_{min} and y^{aug}_{max}: process in the same way as x;
$w^{aug} = (x^{aug}_{min}, y^{aug}_{min}, x^{aug}_{max}, y^{aug}_{max})$;
$L_{w^{aug}} = filter(image(w^{aug}), 'laplacian', \gamma)$;
$L'_{w^{aug}} = norm(resize(|L_{w^{aug}}|, [100, 100]), 1)$;
$L_{max} = max(L'_{w^{aug}})$;
for $i = 1, 2, \ldots, 100$ **do**
  for $j = 1, 2, \ldots, 100$ **do**
    if $L'_{w^{aug}}(i, j) < 0.1 * L_{max}$ **then**
      $L'_{w^{aug}}(i, j) = 0$
    end if
  end for
end for
current centroid: $(x'_c, y'_c) \leftarrow$ average energy point of $L'_{w^{aug}}$;
while energy in $w'' < 0.98 * \sum(L'_{w^{aug}})$ **do**
  $w'' = (x''_{min}, y''_{min}, x''_{max}, y''_{max}) \leftarrow$ update by expanding bounding box in four directions $(-x, -y, x+, y+)$ from the current centroid (x'_c, y'_c).
end while
project w'' into original image: $w' = (x'_{min}, y'_{min}, x'_{max}, y'_{max}) \leftarrow w'' = (x''_{min}, y''_{min}, x''_{max}, y''_{max})$

(resp. small) when detecting very small (resp. large) objects due to the restrictions of the size of the root filter and the scale of the feature pyramid. To improve localization and obtain a more precise prediction of the bounding box aspect ratio, we post-process each bounding box by enlarging or shrinking (*ES* post-processing) it to cover the object as much as possible. This is done using an improved version of the method proposed in [Ke *et al.* 2006], which measures the amount of area that the edge energy occupies. In brief, we first augment the original bounding box $w = (x_{min}, y_{min}, x_{max}, y_{max})$ to 120% of the original width and height (*i.e.*, 144% in total area, denoted as $w^{aug} = (x^{aug}_{min}, y^{aug}_{min}, x^{aug}_{max}, y^{aug}_{max})$. We expand from the centroid if applicable. Otherwise, we stop when reaching the border of the image and calculate the absolute values of the gradients $L_{w^{aug}}$ by applying a 3×3 Laplacian filter with $\gamma = 0.2$ over the augmented bounding box. To simplify calculation of the edge spatial distribution, we then resize the gradient magnitude image size to 100×100 and normalize the image sum to 1, *i.e.*, $L'_{w^{aug}}$. Moreover, we set the values that are less than 10% of the maximum L_{max} to 0. Finally, we expand the bounding box in four directions from the current centroid (x'_c, y'_c) and stop when it contains 98% of the total gradient magnitude (edge energy) in the augmented box. The detailed algorithms are shown in Algorithm 1.

This post-processing technique is not only able to crop out plain background regions, but can also expand to cover the foreground regions which are not encompassed by the original box. However, the cropping method in [Pandey & Lazebnik 2011] can only shrink to reduce the background. Fig. 3.5 shows a few examples of our bounding box post-processing results. It is also worth noticing that this post-processing technique works efficiently for the objects with a unique or plain background, but is of limited help for those with cluttered or textured backgrounds.

3.3 Experimental Evaluation

In this section, we present the experimental results of our proposed framework with two different initialization schemes (*i.e.*, S-WDPMs using single region ini-

Figure 3.5: Examples of bounding box enlarging-and-shrinking. Boxes before (resp. after) post-processing are shown in red (resp. yellow).

tialization and M-WDPMs using multiple region initialization) on the challenging PASCAL VOC 2007 dataset [Everingham *et al.* 2010] and the Microsoft COCO 2014 dataset [Lin *et al.* 2014b].

3.3.1 Experiments with S-WDPMs on PASCAL VOC Subsets

3.3.1.1 Datasets and settings

Following the protocol used in previous works [Pandey & Lazebnik 2011, Deselaers *et al.* 2012, Siva *et al.* 2012, Tang *et al.* 2014a], we evaluate the performance of our proposed S-WDPMs (single region initialized weak DPMs) framework on two subsets from the training and validation set (*trainval*) of the PASCAL VOC 2007 dataset (*VOC07*) [Everingham *et al.* 2010]: *VOC07-6×2* and *VOC07-14*. The *VOC07-6×2* subset contains 6 classes (*aeroplane*, *bicycle*, *boat*, *bus*, *horse* and *motorbike*) with *Left* and *Right* views (aspects) of each class, resulting in a

total of 12 separating classes. The *VOC07-14* subset (same as *PASCAL07-all* defined in [Pandey & Lazebnik 2011]) consists of 42 class/view combinations covering 14 classes and 5 views (*Left*, *Right*, *Frontal*, *Rear* and *Unspecified*). Similar to [Pandey & Lazebnik 2011, Deselaers *et al.* 2012, Siva *et al.* 2012, Tang *et al.* 2014a], we remove all the images annotated as *difficult* or *truncated* in both the training and the evaluation steps.

3.3.1.2 Evaluation protocol

To make fair comparisons with previous works [Pandey & Lazebnik 2011, Deselaers *et al.* 2012, Siva *et al.* 2012, Shi *et al.* 2013], we only choose the detection window with the highest DPMs score per image, although our method can detect multiple instances appearing in the image using the sliding window approach. We also report both results for initial and refined localization as [Pandey & Lazebnik 2011, Siva *et al.* 2012]. A refined localization is obtained by an iteratively trained DPMs detector for one/several iteration(s) to refine the initial detection using the previous annotations as ground truth. Performance is evaluated with the percentage of *training* (train + val) images in which an object is correctly covered by the window (*i.e.* CorLoc [Deselaers *et al.* 2012]), if the strict PASCAL-overlap criterion IoU (intersection-over-union) $\geqslant 0.5$ is satisfied.

3.3.1.3 Experimental evaluation

We compare our S-WDPMs with Weak DPMs [Pandey & Lazebnik 2011], Weak objectness [Deselaers *et al.* 2012] and the Joint topic model [Shi *et al.* 2013]. For the Weak objectness approach [Deselaers *et al.* 2012], the region proposal with the highest "Objectness" score is selected as the predicted window. As shown in Table 3.1, our method outperforms [Deselaers *et al.* 2012] and our baseline approach [Pandey & Lazebnik 2011] on both datasets. Both [Pandey & Lazebnik 2011] and our S-WDPMs use the same HOG feature pyramid for the DPMs. We present our results using two kinds of object proposal generating methods: *Objectness* (*obj*) and *Selective Search* (*SS*). For *obj*, our average performance of initial detec-

Table 3.1: Average localization accuracy (as a %) of our S-WDPMs (single region initialized weak DPMs with HOG features) compared with state-of-the-art competitors on the two variations of the PASCAL VOC 2007 datasets. "crop" and "ES" denote the cropping method from [Pandey & Lazebnik 2011] (denoted as [P&L] in the table) and our enlarging & shrinking post-processing. "*obj*" and "*SS*" denote the objectness and Selective Search region proposal generating method. "*S*" and "*G*" denote the Sampling and Gaussian strategy from [Shi *et al.* 2013] (denoted as [Shi] in the table). Results from [Deselaers *et al.* 2012] are denoted as [Deselaers].

	no post-processing			with post-processing						
	[P&L]	S-WDPMs		[P&L]-crop	S-WDPMs(crop)		S-WDPMs(ES)		[Shi]	
		obj	*SS*		*obj*	*SS*	*obj*	*SS*	*S*	*G*
Dataset	**VOC07-6×2**									
Initialization	37.22	38.74	41.52	44.62	47.85	48.40	48.59	51.01	50.8	**51.5**
Refinement 1	51.63	55.85	63.31	53.11	56.78	64.25	58.02	**67.13**	65.5	66.1
Refinement 2	56.99	59.82	—	59.31	63.31	—	63.91	—	—	—
Refinement 3	59.32	—	—	61.05	—	—	—	—	—	—
[Deselaers]	50.00									
Dataset	**VOC07-14**									
Initialization	19.98	21.73	24.87	23.00	24.20	26.30	25.12	31.84	**32.2**	30.5
Refinement 1	25.11	27.46	31.15	26.38	28.21	33.10	28.94	**34.91**	33.8	32.5
Refinement 2	27.69	28.95	—	29.39	32.87	—	32.82	—	—	—
Refinement 3	28.98	—	—	30.31	—	—	—	—	—	—
[Deselaers]	26.00									

tion before post-processing the bounding boxes on the *VOC07-6×2* and *VOC07-14* subsets is 38.74% and 21.73% respectively, versus 37.22% and 19.98% in [Pandey & Lazebnik 2011]. These improvements are due to the initial object estimate of our method described in Section 3.2.1.4, which ensures better initialization of the root filter of DPMs detectors. We can also observe that both the post-processing method of cropping [Pandey & Lazebnik 2011] (*i.e.*, S-WDPMs(crop) in Table 3.1) and our enlarging-or-shrinking (*i.e.*, S-WDPMs(ES)) post-processing method steadily improves average localization accuracy.

In particular, our ES method is superior to the cropping method of [Pandey & Lazebnik 2011], as our cropped bounding box is able not only to shrink to crop out the background regions, but is also capable of enlarging to cover the

whole foreground object resulting from incomplete coverage of the original window. An example is shown in the last row of Fig. 3.6, where the target object (*motorbike*) is only partially localized by the initial detector (shown in red rectangles in the middle and right images) for both [Pandey & Lazebnik 2011] and our method. However, in the final detection (shown in yellow) after post-processing, our method is able to enlarge the bounding box to approximately include the whole object, while [Pandey & Lazebnik 2011] tends to crop out both foreground and background regions.

Furthermore, the rows starting with "Refinement" in Table 3.1 indicate that localization accuracy can benefit from the iterative refinement process. It is worth mentioning that with a better initialization, our models converge to a steady level of performance after one less round of costly re-training than in [Pandey & Lazebnik 2011] (both using *Objectness*), and achieve slightly better results in the meantime.

The detailed comparisons for our S-WDPMs using *Objectness* with the state-of-the-arts on the *VOC07-6×2* dataset are listed in Table 3.2. The results show that our method outperforms [Pandey & Lazebnik 2011, Siva *et al.* 2012, Deselaers *et al.* 2012] for many of the categories. In particular, our method achieves the state-of-the-art results in the classes where the target object possesses the most salient regions in that category (*e.g.*, *aeroplane, bus, horse*). Interestingly, even without the refinement process, the accuracy of our method in certain categories (*e.g.*, *aeroplane left*) is superior to competitors using the time-consuming refinement procedure. Fig. 3.6 visually compares some of our results with those of [Pandey & Lazebnik 2011]. We also list the co-localization results of [Tang *et al.* 2014a], which does not utilize negative images.

We find that the best detection result using *Selective Search* (63.31%) is 3.49% better than *Objectness* (59.82%) within the same S-WDPMs detection model without post-processing, and is 3.22% better (67.13% *vs.* 63.91%) with post-processing, on the *VOC07-6×2* dataset. This tallies with the conclusion in [Hosang *et al.* 2014], where *Selective Search* provides more reliable detection proposals than *Objectness*. Moreover, it achieves comparable or slightly better results than the sophisticated

Table 3.2: Class level localization accuracy (as a %) for the *VOC07-6×2* dataset for our S-WDPMs(ES) using *Objectness* proposals *vs.* [Pandey & Lazebnik 2011, Deselaers *et al.* 2012, Siva *et al.* 2012, Tang *et al.* 2014a].

	Initialization				Refined by detector		
	ours	[P&L-11]	[Siva-12]	[Tang-14]	**ours**	[P&L-11]	[Deselaer-12]
aero left	**65.1**	55.8	39.1	41.9	**69.7**	65.1	58.0
aero right	**64.1**	61.5	50.0	51.3	**84.6**	82.1	59.0
bike left	31.3	**31.3**	28.4	25.0	85.4	**87.5**	46.0
bike right	42.0	**44.0**	30.6	24.0	54.0	**68.0**	40.0
boat left	9.1	4.6	**15.1**	11.4	**13.6**	2.3	9.0
boat right	9.3	9.3	**20.7**	11.6	14.0	7.0	**16.0**
bus left	23.8	23.8	31.0	**38.1**	**42.9**	28.6	38.0
bus right	**65.2**	52.2	35.1	56.5	69.6	47.8	**74.0**
horse left	**64.6**	60.4	48.5	43.8	**87.5**	83.3	58.0
horse right	**73.9**	67.4	45.2	52.2	76.1	**80.4**	52.0
mbike left	**64.1**	48.7	46.3	51.3	87.2	**92.3**	67.0
mbike right	70.6	**76.5**	55.3	64.7	82.4	**88.2**	76.0
mean	**48.6**	44.6	37.1	39.3	**63.9**	61.1	50.0

joint topic learning models in [Shi *et al.* 2013] when running DPMs refinement only once. As shown in Table 3.1, *SS* also outperforms *obj* on the *VOC07-14* dataset. Consequently, we entirely adopt the *Selective Search* method ("fast" option) for our subsequent experiments.

The localization accuracy on full PASCAL VOC 2007 trainval set and detection precision on test set using S-WDPMs are shown in the first row of Table 3.3 and Table 3.4.

3.3.2 Experiments with M-WDPMs on PASCAL VOC

3.3.2.1 Dataset and settings

We evaluate our generalized model: M-WDPMs (multiple region initialized weak DPMs) on the far more challenging dataset: the whole PASCAL VOC 2007 dataset. This contains a total number of 9963 images of 20 object categories, which are split

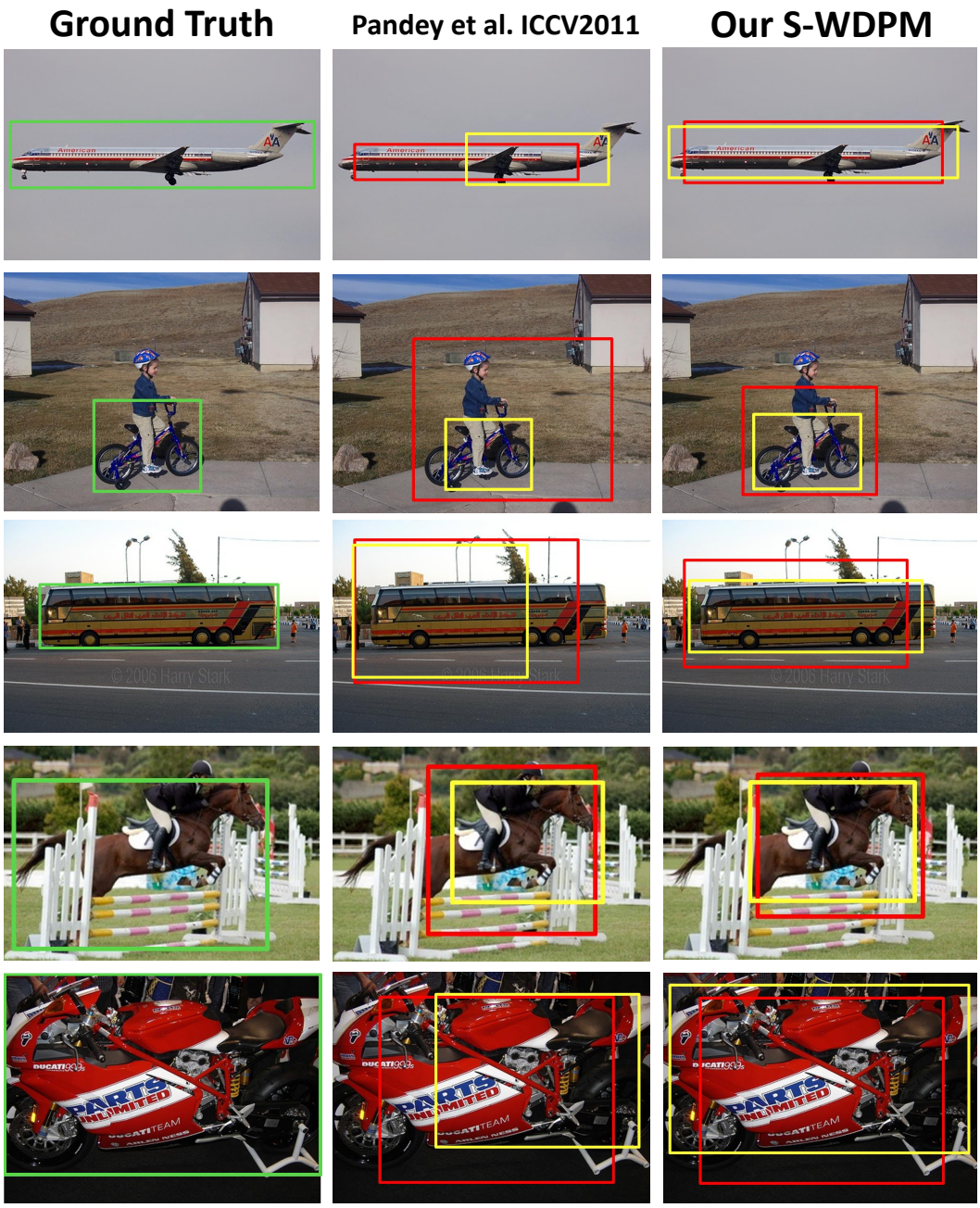

Figure 3.6: Examples of localization results for our S-WDPMs on PASCAL VOC 2007 images. The left column: ground-truth bounding boxes in green rectangles. The middle and right columns are detection results with [Pandey & Lazebnik 2011] and our S-WDPMs framework, respectively. Initial detections are shown in red, while detections refined by detectors are shown in yellow. Both results use the individual post-processing approach.

into training (2501), validation (2510) and test (4952) sets. This dataset is challenging because it has large inter-class similarities, intra-class variances, cluttered backgrounds, and scale changes. We only use the image level category labels for this task. Moreover, images labeled as "difficult" are discarded as common practice in previous studies. With respect to M-WDPMs testing, we only run the DPMs once for efficiency, although iterative detector refinement can steadily improve final performance to a certain extent. Annotation accuracy (*i.e.*, correct localization, CorLoc) on the trainval (training + validation) set and average precision (AP) for detection on the test set are reported. For *DeepPyramid* feature extraction, we use NVIDIA GeForce GTX Titan X GPUs, each with a 12 GB memory, thus allowing us to upsample image pyramids to 1713×1713 as in [Girshick *et al.* 2015] to facilitate detection of small objects.

3.3.2.2 Parameter selection

As discussed in Section 3.2.1.4, we can generate Q region estimations for each image. Q is a parameter which impacting the quality of the positive training examples. If it is too large, there would be an enormous number of noisy samples for latent class learning. However, if it is set to be very small, the instances in the original image could not be comprehensively represented. Therefore, we experimentally vary $Q = \{3, 5, 10, 15, 20, 30\}$ to see which one performs best on the PASCAL VOC 2007 validation set. We implement this by directly measuring average annotation accuracy for all classes, on the generated bounding boxes (Q per image) with the PASCAL-overlap criterion. Figure 3.7 shows annotation accuracy for different Q. We find that $Q = 10$ obtains the best result (36.1% average accuracy). When it is very small (*e.g.*, 3), performance drops dramatically to 26.8%. This is because some of the "good" region proposals are not selected due to very small Q, while some selected "bad" regions may degenerate the model. When Q rises from 10 to 30, performance deteriorates progressively.. One explanation for this might be that many object parts or background regions would be included when Q is large. Hence, we set $Q = 10$ in all of our experiments on PASCAL VOC. Fig. 3.8 shows three example images and their 10 selected regions. The κ in Eq. (3.5), which leverages the

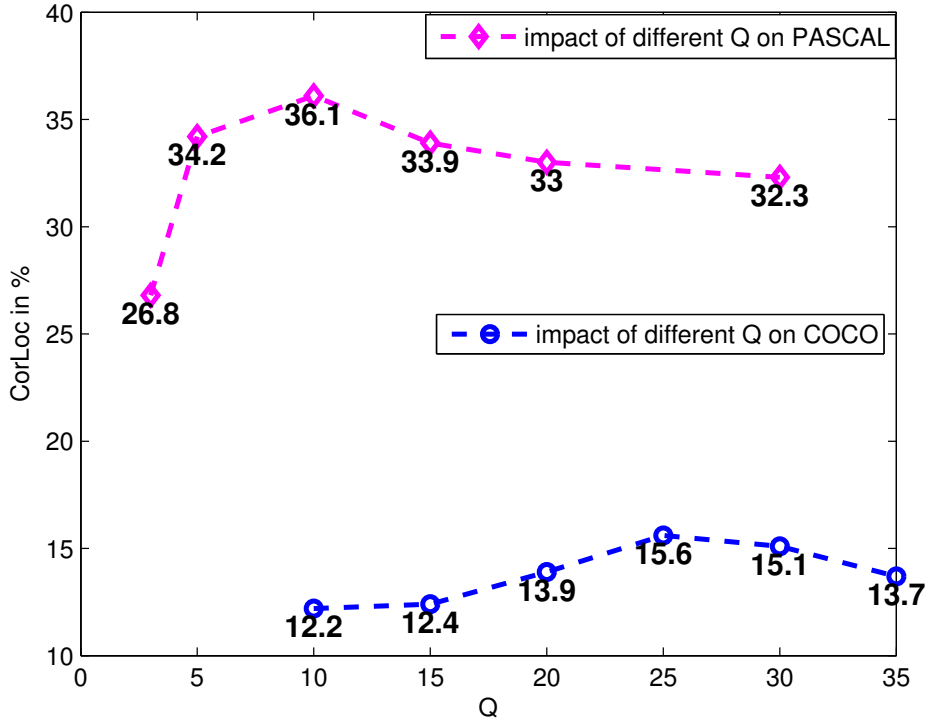

Figure 3.7: The impact of parameter Q (number of selected regions for each image in the multiple region initialization scheme). The average annotation accuracy on PASCAL VOC 2007 validation (HOG feature) and MS COCO 2014 val1 (CNN feature) is evaluated with different Q.

classification and detection scores, is set to 0.7 according to cross-validation on a subset of the validation data.

3.3.2.3 Annotation evaluation

We evaluate the same CorLoc [Deselaers *et al.* 2012] as in Section 3.3.1.2 on the PASCAL VOC 2007 trainval set. Table 3.3 reports our experimental results compared with the state-of-the-art WSL methods for object detection.

Concerning our M-WDPMs-HOG baseline, which computes the HOG features and does not make use of auxiliary training data from the ILSVRC 2012 classification task [Russakovsky *et al.* 2015] as [Wang *et al.* 2015, Bilen *et al.* 2015], it outperforms most of the previous works [Nguyen *et al.* 2009, Siva & Xiang 2011, Shi *et al.* 2012, Shi *et al.* 2013, Siva *et al.* 2012, **?**] (ours: 37.9% *vs.* best from the previ-

Figure 3.8: Three example images and their 10 selected regions (resized to the same squared size for regularity).

ous works (Joint topic): 36.2%). The M-WDPMs-HOG outperforms the S-WDPMs-HOG (by 7.7%) by benefiting initialization of DPMs from multiple regions in the image. Our M-WDPMs-HOG shows modest improvement in most of the classes, thus proving that our multiple region initialization method has very discriminative power for selecting the "good" regions in the original image for training the DPMs root filters.

We also observe that, with the help of auxiliary training data and recently popular deep features, the average accuracy of our M-WDPMs-deep model increases by 4.1% in comparison with the M-WDPMs-HOG model. Moreover, our detection rescoring method (*i.e.*, M-WDPMs-rescore) further improves performance for most

Table 3.3: Comparisons of weakly supervised object detectors on PASCAL VOC 2007 trainval set in terms of correct localization (CorLoc [Deselaers *et al.* 2012], as a %) on positive training images. († indicates methods using auxiliary training data from ILSVRC 2012.)

method / class	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
our S-WDPMs-HOG	49.1	32.8	27.2	9.8	6.6	38.0	46.7	48.2	8.9	35.7
our M-WDPMs-HOG	67.9	52.4	34.4	21.9	12.1	42.0	59.9	58.4	9.9	42.0
our M-WDPMs-deep†	72.0	58.8	38.5	24.6	14.8	46.2	63.4	63.0	18.4	49.9
our M-WDPMs-rescore†	**80.3**	59.1	38.9	**26.0**	14.9	48.8	65.4	65.1	16.6	**58.5**
Joint Learning [Nguyen *et al.* 2009]	30.7	16.5	23.0	14.9	4.9	29.6	26.5	35.3	7.2	23.4
MI-SVM [Andrews *et al.* 2003]	37.8	17.7	26.7	13.8	4.9	34.4	33.7	46.6	5.4	29.8
Model Drift [Siva & Xiang 2011]	42.4	46.5	18.2	8.8	2.9	40.9	73.2	44.8	5.4	30.5
MIL-Negative [Siva *et al.* 2012]	45.8	21.8	30.9	20.4	5.3	37.6	40.8	51.6	7.0	29.8
Transfer Learning [Shi *et al.* 2012]	54.7	22.7	33.7	24.5	4.6	33.9	42.5	57.0	7.3	39.1
Joint Topic [Shi *et al.* 2013]	67.3	54.4	34.3	17.8	1.3	46.6	60.7	**68.9**	2.5	32.4
Convex Cluster.† [Bilen *et al.* 2015]	66.4	59.3	42.7	20.4	**21.3**	**63.4**	**74.3**	59.6	21.1	58.2
LCL-pLSA† [Wang *et al.* 2015]	80.1	**63.9**	**51.5**	14.9	21.0	55.7	74.2	43.5	**26.2**	53.4

method / class	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mean**
our S-WDPMs-HOG	15.3	34.5	42.2	49.5	16.7	13.8	31.6	26.3	47.8	23.1	30.2
our M-WDPMs-HOG	13.5	38.9	48.1	58.6	20.4	19.5	40.8	24.9	48.9	42.7	37.9
our M-WDPMs-deep†	17.0	40.3	52.6	63.2	22.2	22.9	46.1	26.2	52.8	46.8	42.0
our M-WDPMs-rescore†	17.3	42.7	**58.8**	**69.6**	22.8	20.7	52.9	24.0	53.3	46.6	44.1
Joint Learning [Nguyen *et al.* 2009]	20.5	32.1	24.4	33.1	17.2	12.2	20.8	28.8	40.6	7.0	22.4
MI-SVM [Andrews *et al.* 2003]	14.5	32.8	34.8	41.6	19.9	11.4	25.0	23.6	45.2	8.6	25.4
Model Drift [Siva & Xiang 2011]	19.0	34.0	48.8	65.3	8.2	10.6	16.7	32.3	54.8	5.5	30.4
MIL-Negative [Siva *et al.* 2012]	**27.5**	41.3	41.8	47.3	24.1	12.2	28.1	32.8	48.7	9.4	30.2
Transfer Learning [Shi *et al.* 2012]	24.1	43.3	41.3	51.5	**25.3**	13.3	28.0	29.5	54.6	11.8	32.1
Joint Topic [Shi *et al.* 2013]	16.2	**58.9**	51.5	64.6	18.2	3.1	20.9	34.7	**63.4**	5.9	36.2
Convex Cluster.† [Bilen *et al.* 2015]	14.0	38.5	49.5	60.0	19.8	**39.2**	41.7	30.1	50.2	44.1	43.7
LCL-pLSA† [Wang *et al.* 2015]	16.3	56.7	58.3	69.5	14.1	38.3	**58.8**	**47.2**	49.1	**60.9**	**48.5**

of the categories. The average improvement for detection rescoring on all 20 classes is 2.1% (44.1% *vs.* 42.0%). Our M-WDPMs-rescore method is slightly better than the newly invented convex clustering approach [Bilen *et al.* 2015], but is worse than the LCL-pLSA method [Wang *et al.* 2015] on average. Though [Wang *et al.* 2015] achieves state-of-the-art performance on many classes, it depends on more sophisticated Super-Vector (SV) coding [Zhou *et al.* 2010] of the deep CNN features, thus

unfortunately increasing feature dimensionality (*e.g.*, 10,000 visual words). It also fails in some categories such as *boat* and *table*. However, our M-WDPMs-rescore exhibits a steady agreeable performance in all the categories with acceptable feature dimension (256 dimensional *conv 5* features for detection and 4,096 dimensional *fc6* features for classification). In particular, our M-WDPMs-rescore works well in categories where target objects are relatively salient, such as *aeroplane, boat, cow, horse* and *motorbike*. Among these categories, *cow, horse* and *motorbike* have deformable shapes, thus ensuring good detection for the DPM.

3.3.2.4 Detection evaluation

Table 3.4 shows the comparison of our M-WDPMs and other methods for object detection on the PASCAL VOC 2007 test set. Our M-WDPMs-HOG baseline method achieves an mAP of 23.6%, which outperforms [Siva & Xiang 2011] (13.9%) by a large margin, and is slightly better than [Cinbis *et al.* 2014] (22.4%). Both [Siva & Xiang 2011] and [Cinbis *et al.* 2014] represent the image windows with a SIFT [Lowe 1999] descriptor. [Siva & Xiang 2011] uses a Bag-of-Words (BOW) [Csurka *et al.* 2004] histogram with 2000 dimensions, while [Cinbis *et al.* 2014] use Fisher Vectors (FV) encoding [Perronnin *et al.* 2010] to represent the candidate windows. [Pandey & Lazebnik 2011] uses the same HOG pyramid features. M-WDPMs shows consistently better performance than S-WDPMs (19.1%), except for the *sofa* category, where S-WDPMs shows trivial superiority. Among these methods that adopt low level visual features, our M-WDPMs-HOG works best. Although [Song *et al.* 2014a] utilizes powerful deep CNN features to represent the discovered object windows, its performance (22.7%) is more or less the same with our HOG based M-WDPMs, which proves the stronger discrimination of our window selection method. When using the deep features with additional training data from ImageNet [Russakovsky *et al.* 2015], our M-WDPMs-deep can achieve an mAP of 25.7%. The boost (2.1%) is not as much as that of the annotation task (4.1%, see Section. 3.3.2.3), it is probably due to the use of distinct measuring criteria (mean average precision *v.s* percent of correct localization). Our detection rescoring method M-WDPMs-rescore continues to improve the average precision (mAP

Table 3.4: Comparison of weakly supervised object detectors on PASCAL VOC 2007 in terms of AP (Average Precision, as a %) in the test set. († supervised methods using object level annotations.)

method / class	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
our S-WDPMs-HOG	26.2	25.0	8.8	9.1	6.5	37.4	40.7	22.9	5.8	19.8
our M-WDPMs-HOG	34.5	41.6	10.0	14.1	9.0	39.8	43.9	26.6	5.8	22.8
our M-WDPMs-deep	38.3	43.2	18.1	15.9	10.3	40.2	41.9	33.1	6.2	31.4
our M-WDPMs-rescore	43.3	43.5	18.6	**16.8**	10.5	**45.2**	42.3	33.8	6.6	**37.2**
Model Drift [Siva & Xiang 2011]	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3
Multi-fold MIL [Cinbis *et al.* 2014]	35.8	40.6	8.1	7.6	3.1	35.9	41.8	16.8	1.4	23.0
Min-Supervision [Song *et al.* 2014a]	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9
Pattern Config [Song *et al.* 2014b]	36.3	**47.6**	23.3	12.3	11.1	36.0	46.6	25.4	0.7	23.5
Posterior Reg. [Bilen *et al.* 2014]	42.2	43.9	23.1	9.2	**12.5**	44.9	45.1	24.9	8.3	24.0
Convex Cluster. [Bilen *et al.* 2015]	46.2	46.9	**24.1**	16.4	12.2	42.2	**47.1**	35.2	7.8	28.3
LCL-pLSA [Wang *et al.* 2015]	**48.8**	41.0	23.6	12.1	11.1	42.7	40.9	**35.5**	**11.1**	36.6
†DPMs 5.0 [Felzenszwalb *et al.* 2010b]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1
†DP-DPMs conv5 [Girshick *et al.* 2015]	42.3	65.1	32.2	24.4	36.7	56.8	55.7	38.0	28.2	47.3
†R-CNN [Girshick *et al.* 2014]	68.1	72.8	56.8	43.0	36.8	66.3	74.2	67.6	34.4	63.5

method / class	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mean**
our S-WDPMs-HOG	10.6	20.6	27.9	35.1	8.2	6.6	15.3	14.9	27.8	12.2	19.1
our M-WDPMs-HOG	10.8	24.1	32.2	41.7	10.0	12.3	22.5	14.6	32.9	19.1	23.6
our M-WDPMs-deep	11.3	27.4	34.3	45.2	12.7	12.5	25.0	14.9	34.3	19.1	25.7
our M-WDPMs-rescore	12.5	32.7	**36.7**	50.8	14.1	13.8	**28.2**	14.7	**38.0**	20.6	27.7
Model Drift [Siva & Xiang 2011]	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0	13.9
Multi-fold MIL [Cinbis *et al.* 2014]	4.9	14.1	31.9	41.9	**19.3**	11.1	27.6	12.1	31.0	40.6	22.4
Min-Supervision [Song *et al.* 2014a]	10.0	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
Pattern Config [Song *et al.* 2014b]	12.5	23.5	27.9	40.9	14.8	19.2	24.2	17.1	37.7	11.6	24.6
Posterior Reg. [Bilen *et al.* 2014]	13.9	18.6	31.6	43.6	7.6	**20.9**	26.6	20.6	35.9	29.6	26.4
Convex Cluster. [Bilen *et al.* 2015]	12.7	21.5	30.1	42.4	7.8	20.0	26.8	20.8	35.8	29.6	27.7
LCL-pLSA [Wang *et al.* 2015]	**18.4**	**35.3**	34.8	**51.3**	17.2	17.4	26.8	**32.8**	35.1	**45.6**	**30.9**
†DPMs 5.0 [Felzenszwalb *et al.* 2010b]	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
†DP-DPMs conv5 [Girshick *et al.* 2015]	37.1	39.2	61.0	56.4	52.2	26.6	47.0	35.0	51.2	56.1	44.4
†R-CNN [Girshick *et al.* 2014]	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8	58.5

= 27.7%) for nearly all classes except for the *sofa* class (0.2% decrease). Its performance is better when compared with [Song *et al.* 2014b, Bilen *et al.* 2014], and it displays the same range of performance in comparison with [Bilen *et al.* 2015].

The performance gap (3.2%) between our method and that of [Wang *et al.* 2015] might be partly caused by the use different deep feature representations as discussed in Section 3.3.2.3. We conjecture that our detection performance could be further boosted if a complex feature encoding method such as SV [Zhou *et al.* 2010] was adopted as [Wang *et al.* 2015]. We achieve the best detection results for the *boat*, *bus*, *cow*, *horse*, *sheep* and *train* classes for this dataset. We attribute the success on these categories to object saliency (*e.g.*, *boat*, *bus*), deformable structures (*e.g.*, *cow*, *horse*, *sheep*), and possibly their combination (*e.g.*, *train*) which united by our framework. Image saliency and object structures provide good representations for these kinds of object categories. Hence, the combination of the two ensures good detection results on these categories. And our M-WDPMs yields moderately low average precision on categories such as *bird*, *bottle*, *chair* and *potted plant*. These categories typically have notably small and/or textured instances, where object proposal method such as Selective Search can be misleading, and they are hard to detect even by supervised DPMs [Felzenszwalb *et al.* 2010b, Girshick *et al.* 2015].

In addition, we provide the results obtained by popular supervised object detection methods [Felzenszwalb *et al.* 2010b, Girshick *et al.* 2015, Girshick *et al.* 2014] in the bottom lines of Table 3.4. It is clear that there is still a gap between the weakly supervised detection framework and supervised frameworks, although our weakly supervised DPMs yields better results for some classes (*e.g.*, *aeroplane*, *cat*, *dog*, *sheep*) than the supervised DPMs 5.0 [Felzenszwalb *et al.* 2010b] which uses the low level HOG feature. The state-of-the-art supervised object detection framework (*i.e.*, Faster R-CNN [Ren *et al.* 2015a]) achieves 78.9% mAP by adopting very deep neural networks (VGG-16 [Simonyan & Zisserman 2015]).

3.3.2.5 Error analysis

We present an analysis of the types of errors that our M-WDPMs make on the PASCAL VOC 2007 test set. We use the diagnosis tool of [Hoiem *et al.* 2012] and consider four types of false positive (FP) errors: *Loc* (poor localizations), *Sim* (confusion with similar objects), *Oth* (confusion with other objects, *e.g.*, correctly localizing an object but classifying it to a wrong class) and *BG* (confusion with background or

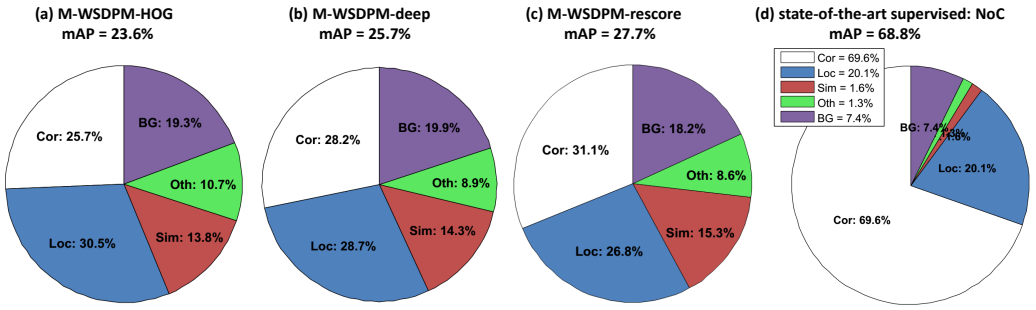

Figure 3.9: Analysis of top-ranked detections on PASCAL VOC 2007 test set. Pie charts show the distributions of the true positives (TP) and false positives (FP) generated by the detection error analysis tool of [Hoiem *et al.* 2012]. Percentage of the top T detections (T is the number of ground truth objects in the whole test dataset) that are correct (Cor), or false positives due to poor localization (Loc), confusion with similar objects (Sim), confusion with other objects (Oth), or confusion with background or unlabeled objects (BG) [Hoiem *et al.* 2012]. The three charts on the left show the analysis of our methods, while the one on the right is the analysis of the state-of-the-art supervised detection results obtained by NoC [Ren *et al.* 2015b].

unlabeled objects). In addition, *Cor* indicates correctly located true positives (TP).

We visually show the fraction of correct detections (TP) and errors of each kind (FP) among the top ranking T windows in Figure 3.9, where T is the number of ground-truth object windows in the PASCAL VOC 2007 test set.

We consider the M-WDPMs-HOG as our baseline and show the distribution of TP and each kind of FP in Figure 3.9(a). We can see that the majority of errors are due to poor localizations (*Loc*) and confusion with background regions (*BG*). When adopting the deep features, our M-WDPMs-deep encounters fewer *Loc* and *Oth*, but continues to suffer from the *Sim* and *BG* error (as shown in Figure 3.9(b)). On the contrary, after detection rescoring, our best performing method M-WDPMs-rescore has fewer errors caused by *Loc*, *BG* and *Oth* (Figure 3.9(c)), thus confirming that our rescoring approach is very efficient in excluding the background regions and avoiding misclassification. Figure 3.9(d) shows the error distribution of the state-of-the-art supervised object detection framework NoC (Networks on Convolutional feature maps) [Ren *et al.* 2015b]. NoC adopt even deeper VGG-16 nets [Simonyan & Zisserman 2015] with bounding box fine-tuning on PASCAL VOC 2007+2012 trainval. A comparison between NoC and our M-WDPMs indicates

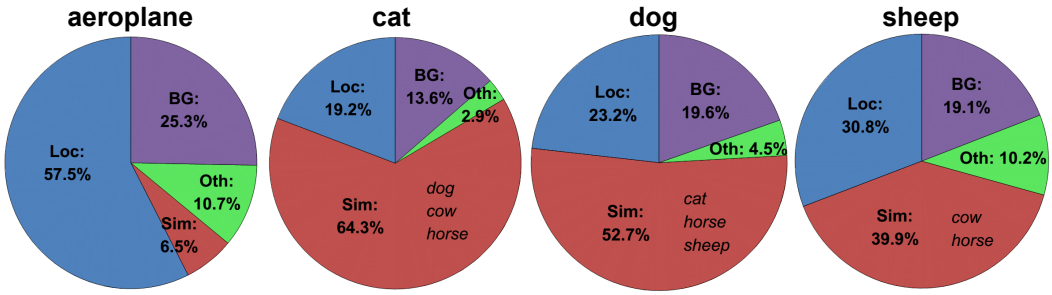

Figure 3.10: Analysis of false positives for some classes on which our M-WDPMs outperforms DPMs 5.0 [Felzenszwalb *et al.* 2010b]. Each category named within "Sim" shows the category names on which detector tends towards confusion.

that: (1) a deeper network helps increase correct localization (*Cor*) substantially; (2) fine-tuning deep CNN and supervised training with ground-truth bounding boxes yield far fewer *Sim* and *Oth* errors.

We also display some class specific false positive analysis in Fig. 3.10, on the classes where our M-WDPMs outperforms DPMs 5.0 [Felzenszwalb *et al.* 2010b].

3.3.2.6 Running time

The time it takes to extract the Selective Search region proposals (can be shared among different detector learning) is 10.27s. Reference region computation takes 778ms, generation of initial object estimations from region proposals and reference regions takes 190ms, while computation of CNN features is 18.97s and *conv5* feature pyramid from CNN feature is 631ms. Running time is averaged on 100 random PASCAL images, and is evaluated on an Intel Core i7-5960X CPU @ 3.00GHz with 32GB memory and a single NVIDIA Titan X GPU. For M-WDPMs, training binary SVM and learning latent class takes 228.20s on the *horse* class and 196.05s on the *motorbike* class (except for CNN pre-training and feature extraction time). Besides, training of the *horse* DPMs detector takes 84.82 minutes and 76.45 minutes for *motorbike*. Running a detector costs 9.76s per image (including rescoring time) on average on the PASCAL dataset.

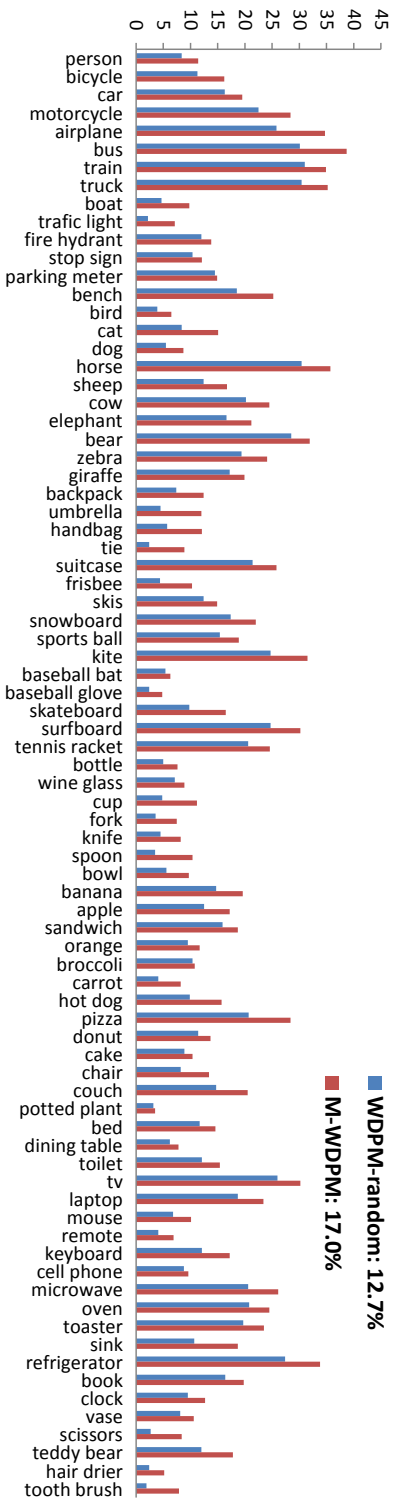

Figure 3.11: Detection results of weakly supervised DPMs detectors on MS COCO 2014 val2 in terms of AP (Average Precision, as a %). For both methods, deep *con5* features are adopted.

3.3.3 Preliminary Results with M-WDPMs on MS COCO

The Microsoft Common Objects in Context (MS COCO) dataset [Lin *et al*. 2014b] involves 80 object categories. It contains considerably more object instances per image (7.7) compared to PASCAL VOC (2.5), and has 82,783 training images and 40,504 validating images in the 2014 release (COCO 2014). We split the validation set equally into val1 and val2, where val1 is used as a validation set and val2 is used as a test set. In spite of this, this subset of MS COCO is much larger and more difficult than PASCAL VOC. We evaluate the PASCAL VOC metric (mAP @IoU = 0.5) on val2.

We set the parameter Q to 25 regions, since there are significantly more object instances per image on MS COCO than on PASCAL VOC. The influence of Q on MS COCO is shown in Figure 3.7, while the rescoring weight κ is set to 0.8 by choosing from [0.5, 1.0] on val1. The increase of κ on MS COCO probably means that there is a larger number of smaller objects in this dataset and that the detector has more influence than the classifier on the final detection score. The other training and testing settings of M-WDPMs remain as the same as on PASCAL VOC. We compare our method with the WDPM-random baseline method [Pandey & Lazebnik 2011], which sets a large random window as initialization. For both of these two methods, we adopt deep *con5* feature pyramids.

Fig. 3.11 shows the detection results of our M-WDPMs and the WDPMs-random baseline. Overall, our M-WDPMs results in 17.0% mAP on this MS COCO val2 set, boosting the mAP by 4.3 points over the WDPMs-random. The results on 20 common categories in MS COCO are significantly lower than on PASCAL. This is because there are far more small objects on COCO, making it a fairly challenging dataset for detection. We observed that our M-WDPMs exhibits a relatively good performance on similar categories both in COCO and in PASCAL, such as *aeroplane, bus, horse, motorbike* and *train*, and has favorable performances on *truck, bear* and *oven*, etc. classes in COCO. This confirms that our M-WDPMs is capable of detecting object categories that are salient visually and/or deformable structurally.

3.4 Summary

In this chapter, we proposed a model enhancing weakly supervised learning by emphasizing the importance of location and size of the initial class specific root filter of deformable part-based models. We follow the general setup of [Pandey & Lazebnik 2011] and introduce several substantial improvements to the weakly supervised deformable part-based model (DPMs). The main contributions included a new selection model based on generic "objectness" (region proposals) and visual saliency to adaptively select a reliable set of candidate windows which tend to represent the object instances in the image, and a latent class learning process by coarsely classifying a candidate window into either a target object or a non-target class. Furthermore, we incorporate the contextual information from image classification, by combining the image-level classification score with object-level DPM detection score, to obtain a final score for detection. We also designed a flexible enlarging-and-shrinking post-processing procedure to modify the output bounding boxes of DPMs, both of which can effectively further improve the final accuracy. Experimental results on the PASCAL VOC 2007 database according to various criteria demonstrate that our proposed framework is efficient and competitive with the state-of-the-art, especially for the object categories which are relatively salient and deformable. We also report some preliminary weakly supervised detection results on the very challenging MS COCO 2014 dataset.

The weakly supervised deformable part-based models have decent performance on mid-level scale datasets such as PASCAL VOC, since DPMs benefit from the relaxed template relation by splitting a single rigid model into smaller part models, and each part model learns more shape details of the object on a finer resolution. However, each fine part template can only handle a specific kind of object deformation or view change since it is sensitive to position, scale, viewpoint, etc.. Hence the complexity becomes intractable for very large scale datasets such as MS COCO and ImageNet, where the object deformation is often very large. In next chapter, we will study the convolutional neural network (CNN) based approach which is more capable in handling unconstrained deformation problems of very

large scale datasets.

Chapter 4

Large Scale Semi-supervised Object Detection Using Visual and Semantic Knowledge Transfer

Contents

4.1 Introduction **82**

4.2 Task Definition **86**

4.3 Similarity-based Knowledge Transfer **86**

4.3.1 Background on LSDA 86

4.3.2 Knowledge Transfer via Visual Similarity 89

4.3.3 Knowledge Transfer via Semantic Relatedness 91

4.3.4 Mixture Transfer Model 93

4.3.5 Transfer on Bounding-box Regression 94

4.4 Experiments **96**

4.4.1 Dataset Overview 96

4.4.2 Implementation Details 96

4.4.3 Quantitative Evaluation on the "Weakly Labeled" Categories with "*Alex-Net*" 97

4.4.4 Experimental Results with "*VGG-Nets*" 104

4.4.5 Experimental Results with Bounding-box Regression 105

4.4.6 Detection Error Analysis 105

4.5 Summary **108**

Deep CNN-based object detection systems have achieved remarkable success on several large-scale object detection benchmarks. However, training such detectors requires a large number of labeled bounding boxes, which are more difficult to obtain than image-level annotations. Previous work addresses this issue by transforming image-level classifiers into object detectors. This is done by modeling the differences between the two on categories with both image-level and bounding box annotations, and transferring this information to convert classifiers to detectors for categories without bounding box annotations. We improve this previous work by incorporating knowledge about object similarities from visual and semantic domains during the transfer process. The intuition behind our proposed method is that visually and semantically similar categories should exhibit more common transferable properties than dissimilar categories, *e.g.* , a better detector would result by transforming the differences between a dog classifier and a dog detector onto the cat class, than would by transforming from the violin class. Experimental results on the challenging ILSVRC2013 detection dataset demonstrate that each of our proposed object similarity based knowledge transfer methods outperforms the baseline methods. We found strong evidence that visual similarity and semantic relatedness are complementary for the task, and when combined notably improve detection, achieving state-of-the-art detection performance in a semi-supervised setting.

4.1 Introduction

The recent success of deep convolutional neural networks (CNN) [Krizhevsky *et al.* 2012] for object detection, such as DetectorNet [Szegedy *et al.* 2013], OverFeat [Sermanet *et al.* 2014], R-CNN [Girshick *et al.* 2014], SPP-net [He *et al.* 2015], Fast R-CNN [Girshick 2015], Faster R-CNN [Ren *et al.* 2015a], YOLO [Redmon *et al.* 2016] and SSD [Liu *et al.* 2016], is heavily dependent on a large amount of training data manually labeled with object

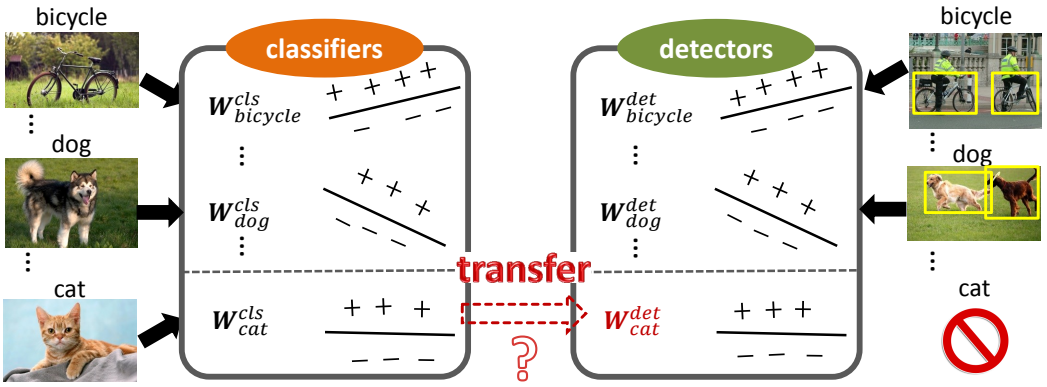

Figure 4.1: In this work, we consider a dataset containing image-level labels for all the categories, while object-level bounding box annotations are only available for some of the categories (*i.e.* weakly labeled categories). How can we transform a CNN classification network into a detection network to detect the weakly labeled categories (*e.g.* , *cat* class)?

localizations (*e.g.* , PASCAL VOC [Everingham *et al.* 2010], ILSVRC (subset of ImageNet) [Russakovsky *et al.* 2015], and Microsoft COCO [Lin *et al.* 2014b] datasets).

Although localized object annotations are extremely valuable, the process of manually annotating object bounding boxes is extremely laborious and unreliable, especially for large-scale databases. On the other hand, it is usually much easier to obtain annotations at *image* level (*e.g.* , from user-generated tags on Flickr or Web queries). For example, ILSVRC contains image-level annotations for 1,000 categories, while object-level annotations are currently restricted to only 200 categories. One could apply image-level classifiers directly to detect object categories, but this will result in a poor performance as there are differences in the statistical distribution between the training data (whole images) and the test data (localized object instances). Previous work by Hoffman *et al.* [Hoffman *et al.* 2014] addresses this issue, by learning a transformation between classifiers and detectors of object categories with *both* image-level and object-level annotations ("strong" categories), and applying the transformation to adapt image-level classifiers to object detectors for categories with *only* image-level labels ("weak" categories). Part of this work involves transferring *category-specific* classifier and detector differences of visually

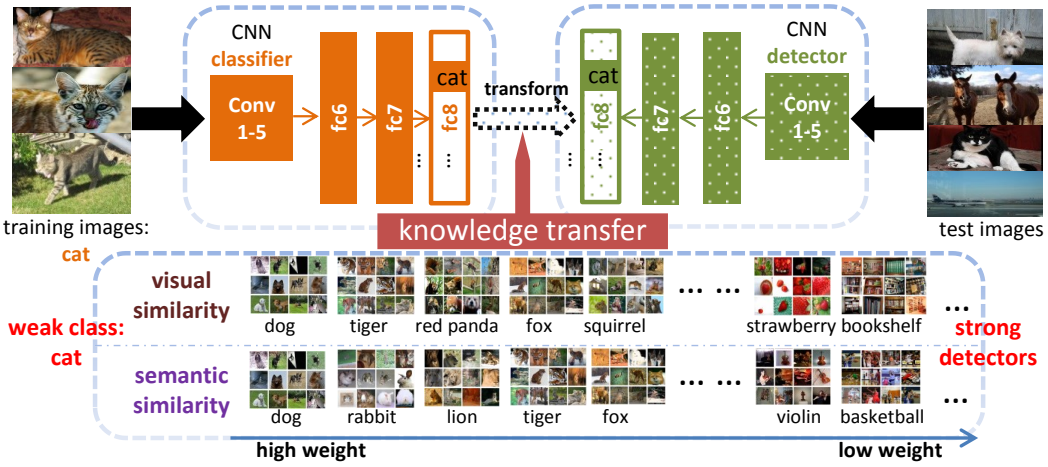

Figure 4.2: An illustration of our similarity-based knowledge transfer model. The question we investigate is whether knowledge about object similarities – visual and semantic – can be exploited to improve detectors trained in a semi-supervised manner. More specifically, to adapt the image-level classifier (up-left) of a "weakly labeled" category (no bounding boxes) into a detector (up-right), we transfer information about the classifier and detector differences of "strong" categories (with image-level and bounding box annotations, bottom of the figure) by favoring categories that are more similar to the target category (*e.g.* , transfer information from *dog* and *tiger* rather than *basketball* or *bookshelf* to produce a *cat* detector).

similar "strong" categories equally to a classifier of a "weak" category to form a detector for that category (Figure 4.1). We argue that more can potentially be exploited from such similarities in an informed manner to improve detection beyond using the measures solely for nearest neighbor selection (see Section 4.3.1). Moreover, since there exists evidence that deep CNNs trained for image classification also learn proxies to objects and object parts [Zhou *et al.* 2015], the transformation from CNN classifiers to detectors is reasonable and practicable.

Our main contribution in this chapter is therefore to incorporate external knowledge about object similarities from visual *and* semantic domains in modeling the aforementioned category-specific differences, and subsequently transferring this knowledge for adapting an image classifier to an object detector for a "weak" category. Our proposed method is motivated by the following observations: (i) category specific difference exists between a classifier and a detector [Girshick *et al.* 2014, Hoffman *et al.* 2014]; (ii) visually and semantically similar

categories may exhibit more common transferable properties than visually or semantically dissimilar categories; (iii) visual similarity and semantic relatedness are shown to be correlated, especially when measured against object instances cropped out from images (thus discarding background clutter) [Deselaers & Ferrari 2011]. Intuitively, we would prefer to adapt a *cat* classifier to a *cat* detector by using the category-specific differences between the classifier and the detector of a *dog* rather than of a *violin* or a *strawberry* (Figure 4.2). The main advantage of our proposed method is that knowledge about object similarities can be obtained without requiring further object-level annotations, for example from existing image databases, text corpora and external knowledge bases.

Our work aims to answer the question: can knowledge about visual and semantic similarities of object categories (and the combination of both) help improve the performance of detectors trained in a weakly supervised setting (*i.e.* by converting an image classifier into an object detector for categories with only image-level annotations)? Our claim is that by exploiting knowledge about objects that are visually and semantically similar, we can better model the category-specific differences between an image classifier and an object detector and hence improve detection performance, without requiring bounding box annotations. We also hypothesize that the combination of both visual and semantic similarities can help further improve the detector performance. Experimental results on the challenging ILSVRC 2013 dataset [Russakovsky *et al.* 2015] validate these claims, showing the effectiveness of our approach of transferring knowledge about object similarities from both visual and semantic domains to adapt image classifiers into object detectors in a semi-supervised manner.

The rest of this chapter is organized as follows. We define the semi-supervised object detection problem in Section 4.2. In Section 4.3, we first review the LSDA framework, then we introduce our two knowledge transferring methods (*i.e.* visual similarity based method and semantic similarity based method) which improve upon LSDA. We present our experimental results and comparisons in Section 4.4. In Section 4.5, we conclude and describe future direction.

4.2 Task Definition

In our semi-supervised learning case, we assume that we have a set of "fully labeled" categories and "weakly labeled" categories. For the "fully labeled" categories, a large number of training images with both image-level labels and bounding box annotations are available for learning the object detectors. For each of the "weakly labeled" categories, we have many training images containing the target object, but we do not have access to the exact locations of the objects. This is different from the semi-supervised learning proposed in previous work [Misra *et al.* 2015, Rosenberg *et al.* 2005, Yang *et al.* 2013], where typically a small amount of fully labeled data with a large amount of weakly labeled (or unlabeled) data are provided for each category. In our semi-supervised object detection scenario, the objective is to transfer the trained image classifiers into object detectors on the "weakly labeled" categories.

4.3 Similarity-based Knowledge Transfer

We first describe the Large Scale Detection through Adaptation (LSDA) framework [Hoffman *et al.* 2014], upon which our proposed approach is based (Section 4.3.1). We then describe our proposed knowledge transfer models with the aim of improving LSDA. Two knowledge domains are explored: (i) visual similarity (Section 4.3.2); (ii) semantic relatedness (Section 4.3.3). Next, we combine both models to obtain our mixture transfer model, as presented in Section 4.3.4. Finally, we propose to transfer the knowledge to bounding-box regression from fully labeled categories to weakly labeled categories in Section 4.3.5.

4.3.1 Background on LSDA

Let $\mathcal{D}$ be the dataset of K categories to be detected. One has access to both image-level and bounding box annotations only for a set of m ($m \ll K$) "fully labeled" categories, denoted as $\mathcal{B}$, but only image-level annotations for the rest of the categories, namely "weakly labeled" categories, denoted as $\mathcal{A}$. Hence, a set of K image

classifiers can be trained on the whole dataset $\mathcal{D}$ ($\mathcal{D} = \mathcal{A} \cup \mathcal{B}$), but only m object detectors (from $\mathcal{B}$) can be learned according to the availability of bounding box annotations. The LSDA algorithm learns to convert $(K - m)$ image classifiers (from $\mathcal{A}$) into their corresponding object detectors through the following steps:

Pre-training: First, an 8-layer (5 convolutional layers and 3 fully-connected (fc) layers) *Alex-Net* [Krizhevsky *et al.* 2012] CNN is pre-trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 classification dataset [Russakovsky *et al.* 2015], which contains 1.2 million images of 1,000 categories.

Fine-tuning for classification: The final weight layer (1,000 linear classifiers) of the pre-trained CNN is then replaced with K linear classifiers. This weight layer is randomly initialized and the whole CNN is then fine-tuned on the dataset $\mathcal{D}$. This produces a classification network that can classify K categories (*i.e.*, K-way softmax classifier), given an image or an image region as input.

Category-invariant adaptation: Next, the classification network is fine-tuned into a detector with bounding boxes of $\mathcal{B}$ as input, using the R-CNN [Girshick *et al.* 2014] framework. As in R-CNN, a background class ($fc8_{\mathcal{BG}}$) is added to the output layer and fine-tuned using bounding boxes from a region proposal algorithm, *e.g.* , Selective Search [Uijlings *et al.* 2013]. The $fc8$ layer parameters are category *specific*, with 4,097 weights ($fc7$ output: 4,096, plus a bias term) in each category, while the parameters of layers 1-7 are category *invariant*. Note that object detectors are not able to be directly trained on $\mathcal{A}$, since the fine-tuning and training process requires bounding box annotations. Therefore, at this point, the category specific output layer $fc8_{\mathcal{A}}$ stays unchanged. The variation matrix of $fc8_{\mathcal{B}}$ after fine-tuning is denoted as $\Delta_{\mathcal{B}}$.

Category-specific adaptation: Finally, each classifier of categories $j \in \mathcal{A}$ is adapted into a corresponding detector by learning a category-specific transformation of the model parameters. This is based on the assumption that the difference between classification and detection of a target object category has a positive correlation with those of similar (close) categories. The transformation is computed by adding a bias vector to the weights of $fc8_{\mathcal{A}}$. This bias vector for category j is measured by the average weight change of its k nearest neighbor categories in set $\mathcal{B}$, from

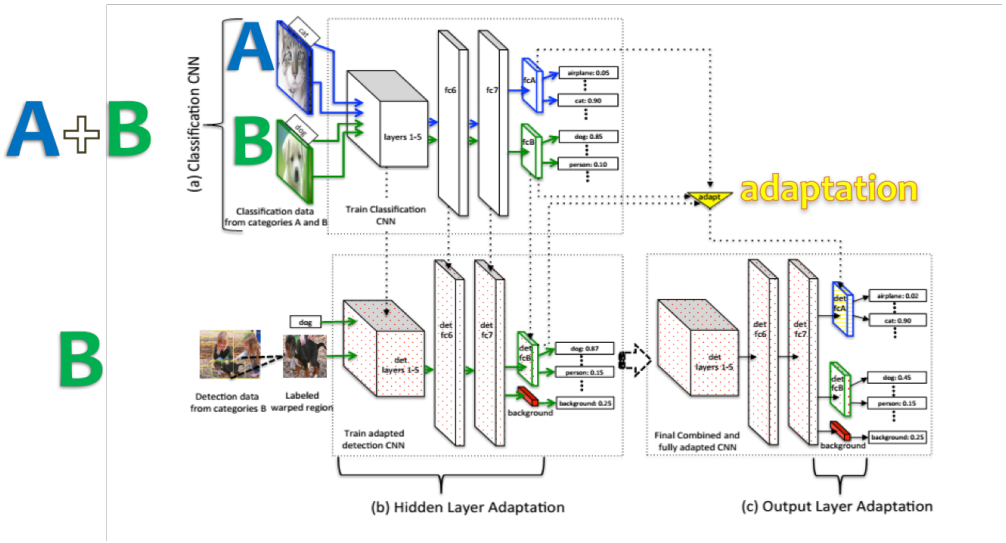

Figure 4.3: The pipeline of the LSDA [Hoffman *et al.* 2014] framework.

classification to detection.

$$\forall j \in \mathcal{A} : \overrightarrow{w_j^d} = \overrightarrow{w_j^c} + \frac{1}{k}\sum_{i=1}^{k} \Delta_{\mathcal{B}_i^j} \tag{4.1}$$

where $\Delta_{\mathcal{B}_i^j}$ is the $fc8$ weight variation of the i^{th} nearest neighbor category in set $\mathcal{B}$ for category $j \in \mathcal{A}$. $\overrightarrow{w^c}$ and $\overrightarrow{w^d}$ are, respectively, $fc8$ layer weights for the fine-tuned classification and the adapted detection network. The nearest neighbor categories are defined as those with nearest L_2-*norm* (Euclidean distance) of $fc8$ weights in set $\mathcal{B}$.

The pipeline of the LSDA framework is shown in Figure 4.3. The fully adapted network is able to detect all K categories in test images. In contrast to R-CNN, which trains SVM classifiers on the output of the $fc7$ layer followed by bounding box regression on the extracted features from the $pool5$ layer of all region proposals, LSDA directly outputs the score of the softmax "detector", and subtracts the background score from this as the final score. This results in a small drop in performance, but enables direct adaptation from a classification network into a detection network on the "weakly labeled" categories, and significantly reduces the training time.

Hoffman *et al.* [Hoffman *et al.* 2014] demonstrated that the adapted model yielded a 50% relative mAP (mean average precision) boost for detection over the classification-only framework on the "weakly labeled" categories of the ILSVRC2013 detection dataset (from 10.31% to 16.15%). They also showed that category-specific adaptation (final LSDA step) contributes least to the performance improvement (16.15% with *vs.* 15.85% without this step), with the other features (adapted layers 1-7 and background class) being more important. However, we found that by properly adapting this layer, a significant boost in performance can be achieved: an mAP of 22.03% can be obtained by replacing the semi-supervised $fc8_{\mathcal{A}}$ weights with their corresponding supervised network weights and leaving the other parameters fixed. Thus, we believe that adapting this layer in an informed manner, such as making better use of knowledge about object similarities, will help improve detection.

In the next subsections, we will introduce our knowledge transfer methods using two different kinds of similarity measurements to select the nearest categories and weight them accordingly to better adapt the $fc8$ layer, which can efficiently convert an image classifier into an object detector for a "weakly labeled" category.

4.3.2 Knowledge Transfer via Visual Similarity

Intuitively, the object detector of an object category may be more similar to those of visually similar categories than of visually distinct categories. For example, a cat detector may approximate a dog detector better than a strawberry detector, since cat and dog are both mammals sharing common attributes in terms of shape (both have four legs, two ears, two eyes, one tail) and texture (both have fur). Therefore, given a "fully labeled" dataset $\mathcal{B}$ and a "weakly labeled" dataset $\mathcal{A}$, our objective is to model the visual similarity between each category $j \in \mathcal{A}$ and all the other categories in $\mathcal{B}$, and to transfer this knowledge for transforming classifiers into detectors for $\mathcal{A}$.

Visual similarity measure: Visual similarity measurements are often obtained by computing the distance between feature distributions such as the $fc6$ or $fc7$ output of a CNN, or in the case of LSDA the $fc8$ layer parameters. In our

work, we instead forward propagate an image through the whole fine-tuned classification network (created by the second step in Section 4.3.1) to obtain a K-dimensional classification score vector. This score vector encodes the probabilities of an image being each of the K object categories. Consequently, for all the positive images of an object category $j \in \mathcal{A}$, we can directly accumulate the scores of each dimension, on a balanced validation dataset. We assume that the normalized accumulated scores (range [0,1]) imply the similarities between category j and other categories: the larger the score, the more it visually resembles category j. This assumption is supported by the analysis of deep CNNs [Agrawal *et al.* 2014, Jia *et al.* 2014, Zeiler & Fergus 2014]: CNNs are apt to confuse visually similar categories, on which they might have higher prediction scores. The visual similarity (denoted s_v) between a "weakly labeled" category $j \in \mathcal{A}$ and a "fully labeled" category $i \in \mathcal{B}$ is defined as:

$$s_v(j,i) \propto \frac{1}{N}\sum_{n=1}^{N} CNN(I_n)_i \tag{4.2}$$

where I_n is a positive image from category j of the validation set of $\mathcal{A}$, N is the number of positive images for this category, and $CNN(I_n)_i$ is the i^{th} CNN output of the softmax layer on I_n, namely, the probability of I_n being category $i \in \mathcal{B}$ as predicted by the fine-tuned classification network. $s_v(j,i) \in [0,1]$ is the degree of similarity after normalization on all the categories in $\mathcal{B}$.

Note that we adopt the $fc8$ outputs since most of the computation is integrated into the end-to-end *Alex-Net* framework except for the accumulation of classification scores in the end, saving the extra effort otherwise required for distance computation if $fc6$ or $fc7$ were to be used (two methods produce similar range of results).

Weighted nearest neighbor scheme: Using Eq. (4.1), we can transfer the model parameters based on a category's k nearest neighbor categories selected by Eq. (4.2). This allows us to directly compare our visual similarity measure to that of LSDA which uses the Euclidean distance between the $fc8$ parameters. An alternative to Eq. (4.1) is to consider a *weighted* nearest neighbor scheme, where weights can be

assigned to different categories based on how visually similar they are to the target object category. This is intuitive, as different categories will have varied degrees of similarity to a particular class, and some categories may have only a few (or many) visually similar classes. Thus, we modify Eq. (4.1) and define the transformation via visual similarity based on the proposed weighted nearest neighbor scheme as:

$$\forall j \in \mathcal{A} : \overrightarrow{w^d_{j\,v}} = \overrightarrow{w^c_j} + \sum_{i=1}^{m} s_v(j,i)\Delta_{\mathcal{B}^j_i} \tag{4.3}$$

It is worth noting that Eq. (4.1) is a special case of Eq. (4.3), where $m = k$ and $s_v(j,i) = 1/k$.

4.3.3 Knowledge Transfer via Semantic Relatedness

Following prior work [Deselaers & Ferrari 2011, Rochan & Wang 2015, Rohrbach *et al.* 2010], we observe that visual similarity is correlated with semantic relatedness. According to [Deselaers & Ferrari 2011], this relationship is particularly strong when measurements are focused on the category instances themselves, ignoring image backgrounds. This observation is quite intriguing for object detection, where the main focus is on the target objects themselves. Hence, we draw on this fact and propose transferring knowledge from the natural language domain to help improve semi-supervised object detection.

Semantic similarity measure: Semantic similarity is a well-explored area within the Natural Language Processing community. Recent advances in word embeddings trained on large-scale text corpora [Mikolov *et al.* 2013a, Pennington *et al.* 2014] have helped progress research in this area, as it has been observed that semantically related word vectors tend to be close in the embedding space, and that the embeddings capture various linguistic regularities [Mikolov *et al.* 2013b]. Thus, we encode each of the K categories as a word vector, more specifically a 300-dimensional word2vec embedding [Mikolov *et al.* 2013a]. As each category is a WordNet [Fellbaum 1998] synset, we represent each category as the sum of the word vectors for each term in its synset, normalized to unit vector by its L_2-*norm*. Out-of-vocabulary words are ad-

dressed by attempting to match case variants of the words (lowercase, Capitalized), *e.g.* , "aeroplane" is not in the vocabulary, but "Aeroplane" is. Failing that, we represent multiword phrases by the sum of the word vectors of each in-vocabulary word of the phrase, normalized to unit vector ("baby"+"bed" for *baby bed*). In several cases, we also augment synset terms with any category label defined in ILSVRC2013 that is not among the synset terms defined in WordNet (e.g. "bookshelf" for the WordNet synset *bookcase*, and "tv" and "monitor" for *display*).

Word embeddings often conflate multiple senses of a word into a single vector, leading to an issue with polysemous words. We observed this with many categories, for example *seal* (animal) is close to *nail* and *tie* (which, to further complicate matters, is actually meant to refer to its clothing sense); or the stationery *ruler* being related to *lion*. Since ILSVRC2013 categories are actually WordNet synsets, it makes perfect sense to exploit WordNet to help disambiguate the word senses. Thus, we integrate corpus-based representations with semantic knowledge from WordNet, by using AutoExtend [Rothe & Schütze 2015] to encode the categories as *synset embeddings* in the original word2vec embedding space. AutoExtend exploits the interrelations between synsets, words and lexemes to learn an auto-encoder based on these constraints, as well as constraints on WordNet relations such as hypernyms (encouraging *poodle* and *dog* to have similar embeddings). We observed that AutoExtend has indeed helped form better semantic relations between the desired categories: *seal* is now clustered with other animal categories like *whale* and *turtle*, and the nearest neighbors for *ruler* are now *rubber eraser*, *power drill* and *pencil box*. In our detection experiments (Section 4.4), we found that while the 'naive' word embeddings performed better than the baselines, the synset embeddings yielded even better results. Thus, we only report the results for the latter.

We represent each category $j \in \mathcal{A}$ and $i \in \mathcal{B}$ with their synset embeddings, and compute the L_2*-norm* of each pair $d_s(j, i)$ as their semantic distance. The semantic similarity $s_s(j, i)$ is inversely proportional to $d_s(j, i)$. We can then transfer the semantic knowledge to the appearance model using Eq. (4.3) or its special case Eq. (4.1) as before.

As our semantic representations are in the form of vectors, we explore an alter-

native similarity measure as used in [Rochan & Wang 2015]. We assume that each vector of a "weakly labeled" category $j \in \mathcal{A}$ (denoted as v_j) can be approximately represented by a linear combination of all the m word vectors in $\mathcal{B}$: $v_j \approx \Gamma_j V$, where $V = [v_1; v_2; \ldots; v_i; \ldots; v_m]$, and $\Gamma_j = [\gamma_j^1, \gamma_j^2, \ldots, \gamma_j^i, \ldots, \gamma_j^m]$ is a set of coefficients of the linear combination. We are motivated to find the solution $\Gamma_j^\star$ which contains as few non-zero components as possible, since we tend to reconstruct category j with fewer categories from $\mathcal{B}$ (sparse representation). This optimal solution $\Gamma_j^\star$ can be formulated as the following optimization:

$$\Gamma_j^\star = \arg\min_{\Gamma_j > 0} (\|v_j - \Gamma_j V\|_2 + \lambda \|\Gamma_j\|_0) \tag{4.4}$$

Note that $\Gamma_j > 0$ is a positive constraint on the coefficients, since negative components of sparse solutions for semantic transferring are meaningless: we only care about the most similar categories and not dissimilar categories. We solve Eq. (4.4) by using the positive constraint matching pursuit (PCMP) algorithm [Gao *et al.* 2012]. Therefore, the final transformation via semantic transferring is formulated as:

$$\forall j \in \mathcal{A} : \overrightarrow{w_{j\,s}^d} = \overrightarrow{w_j^c} + \sum_{i=1}^{m} s_s(j,i) \Delta_{\mathcal{B}_i^j} \tag{4.5}$$

where $s_s(j,i) = \gamma_j^i$ in the sparse representation case.

4.3.4 Mixture Transfer Model

We have proposed two different knowledge transfer models. Each of them can be integrated into the LSDA framework independently. In addition, since we consider the visual similarity at the whole image level and the semantic relatedness at object level, they can be combined simultaneously to provide complementary information. We use a simple but very effective combination of the two knowledge transfer models as our final mixture transfer model. Our mixture model is a linear combination of the visual similarity and the semantic similarity:

$$s = intersect[\alpha s_v + (1-\alpha) s_s] \tag{4.6}$$

where $intersect[\cdot]$ is a function that takes the intersection of cooccurring categories between visual and sparse semantic related categories. $\alpha \in [0, 1]$ is a parameter used to control the relative influence of the two similarity measurements. α is set to 1 when only considering visual similarity transfer, and 0 for the semantic similarity transfer. We will analyze this parameter in Section 4.4.3.

4.3.5 Transfer on Bounding-box Regression

The detection windows generated by the region based detection models are the highest scoring proposals (*e.g.* , Selective Search). In order to improve localization performance, a bounding-box regression stage [Girshick *et al.* 2014] is commonly adopted to post-process the detection windows. This process needs bounding box annotations in training the regressors, which is an obstacle for "weakly labeled" categories in our case. Hence, we propose to transfer the class-specific regressors from "fully labeled" categories to "weakly labeled" categories based on the aforementioned similarity measures.

To train a regressor for each "fully labeled" category, we select a set of N training pairs $\{(\vec{P}^i, \vec{G}^i)\}_{i=1,\ldots,N}$, where $\vec{P}^i = (P^i_x, P^i_y, P^i_w, P^i_h)$ is a vector indicating the center coordinates (P^i_x, P^i_y) of proposal P^i together with P^i's width and height (P^i_w, P^i_h). $\vec{G}^i = (G^i_x, G^i_y, G^i_w, G^i_h)$ is the corresponding ground-truth bounding box. We omit the superscript i except as hereinafter provided. The goal is to learn a mapping function $f(P) = (f_x(P), f_y(P), f_w(P), f_h(P))$ which maps a region proposal P to a ground-truth window G. Each function within $f(P)$ is modeled as a linear function of the $pool5$ features: $f(P) = \mathbf{w}_*^T F_5(P)$, where $\mathbf{w}_*$ is a vector of learnable parameters, $F_5(P)$ is the $pool5$ feature of region proposal P. $\mathbf{w}_*$ can be learned by optimizing the following least squares objective function:

$$\mathbf{w}_* = \arg\min_{\hat{\mathbf{w}}_*} \sum_{i=1}^{N} (\hat{\mathbf{w}}_*^T F_5(P^i) - t^i_*)^2 + \lambda_0 \|\hat{\mathbf{w}}_*\|^2 \tag{4.7}$$

where $t_* = (t_x, t_y, t_w, t_h)$ is the regression target for the training pair (P, G) which

is defined as:

$$\begin{aligned} t_x &= (G_x - P_x)/P_w, \\ t_y &= (G_y - P_y)/P_h, \\ t_w &= \log(G_w/P_w), \\ t_h &= \log(G_h/P_h). \end{aligned} \tag{4.8}$$

The first two specify a scale-invariant translation of the center of the bounding box, while the second two specify log-space translation of the width and height of the bounding box. After learning the parameters of the transformation function, a detection window (region proposal) P can be transformed into a new prediction $\hat{P} = (\hat{P}_x, \hat{P}_y, \hat{P}_w, \hat{P}_h)$ by applying:

$$\begin{aligned} \hat{P}_x &= P_x + P_w f_x(P), \\ \hat{P}_y &= P_y + P_h f_h(P), \\ \hat{P}_w &= P_w \exp(f_w(P)), \\ \hat{P}_h &= P_h \exp(f_h(P)). \end{aligned} \tag{4.9}$$

The training pair (P, G) is selected when the proposal P has maximum IoU overlap with ground-truth bounding box G. The pair (P, G) is discarded if the maximum IoU overlap is less than a threshold (which is set to be 0.6 using a validation set).

For a "weakly labeled" category j, the transformation function can not been explicitly learned due to the absence of ground-truth bounding boxes. However, we can still transfer this knowledge from similar categories in "fully labeled" subset $\mathcal{B}$:

$$\forall j \in \mathcal{A} : \mathbf{w}_j = \sum_{i=1}^{m} s_* \mathbf{w}_i \tag{4.10}$$

where s_* indicates any one of the aforementioned similarity measures.

4.4 Experiments

4.4.1 Dataset Overview

We investigate the proposed knowledge transfer models for large scale semi-supervised object detection on the ILSVRC2013 detection dataset covering 200 object categories. The training set is not exhaustively annotated because of its sheer size. There are also fewer annotated objects per training image than the validation and testing image (on average 1.53 objects for training *vs.* 2.5 objects for validation set). We follow all the experiment settings as in [Hoffman *et al.* 2014], and simulate having access to image-level annotations for all 200 categories and bounding box annotations only for the first 100 categories (alphabetical order). We separate the dataset into classification and detection sets. For the classification data, we use 200,000 images in total from all 200 categories of the training subset (around 1,000 images per category) and their image-level labels. The validation set is roughly split in half: val1 and val2 as in [Girshick *et al.* 2014]. For the detection training set, we take the images with their bounding boxes from only the first 100 categories ($\mathcal{B}$) in val1 (around 5,000 images in total). Since the validation dataset is relatively small, we then augment val1 with 1,000 bounding box annotated images per class from the training set (following the same protocol of [Girshick *et al.* 2014, Hoffman *et al.* 2014]). Finally, we evaluate our knowledge transfer framework on the val2 dataset (9,917 images in total).

4.4.2 Implementation Details

In all the experiments, we consider LSDA [Hoffman *et al.* 2014] as our baseline model and follow their main settings. Following [Hoffman *et al.* 2014], we first use the Caffe [Jia *et al.* 2014] implementation of the "AlexNet" CNN. A pre-trained CNN on ILSVRC 2012 dataset is then fine-tuned on the classification training dataset (see Section. 4.4.1). This CNN is then fine-tuned again for detection on the labeled region proposals of the first 100 categories (subset $\mathcal{B}$) of val1. Selective Search [Uijlings *et al.* 2013] with "fast" mode is adopted to generate the region proposals from all the images in val1 and val2. We also report results using two deeper

models of "VGG-Nets" [Simonyan & Zisserman 2015], namely, the 16-layer model (VGG-16) and the 19-layer model (VGG-19), with the Caffe toolbox. For the semantic representation, we use word2vec CBoW embeddings pre-trained on part of the Google News dataset comprising about 100 billion words [Mikolov *et al.* 2013a]. We train AutoExtend [Rothe & Schütze 2015] using WordNet 3.0 to obtain synset embeddings, and using equal weights for the synset, lexeme and WordNet relation constraints ($\alpha = \beta = 0.33$). As all categories are nouns, we use only hypernyms as the WordNet relation constraint. For the sparse representation of a target word vector in Eq. (4.4), we limit the maximum number of non-zero components to 20, since a target category has strong correlation with a small number of source categories. We set $\lambda = 100$ in Eq. (4.4) and $\lambda_0 = 1000$ in Eq. (4.7) based on a validation set. The other detailed information regarding training and detection can be found in Section 4.3.1.

4.4.3 Quantitative Evaluation on the "Weakly Labeled" Categories with "*Alex-Net*"

Setting LSDA as the baseline, we compare the detection performance of our proposed knowledge transfer methods against LSDA. The results are summarized in Table 4.1. As we are concerned with the detection of the "weakly labeled" categories, we focus mainly on the second column of the table (mean average precision (mAP) on $\mathcal{A}$). Rows 1-5 in Table 4.1 are the baseline results for LSDA. The first row shows the detection results by applying a classification network (*i.e.*, weakly supervised learning, and without adaptation) trained with only classification data, achieving only an mAP of 10.31% on the "weakly labeled" 100 categories. The last row shows the results of an oracle detection network which assumes that bounding boxes for all 200 categories are available (*i.e.*, supervised learning). This is treated as the upper bound (26.25%) of the fully supervised framework. We observed that the best result obtained by LSDA is to adapt both category independent and category specific layers, and transforming with the weighted $fc8$ layer weight change of its 100 nearest neighbor categories (**weighted-100** with 16.33% in Table 4.1). Our

Table 4.1: Detection mean average precision (mAP) on ILSVRC2013 val2. The first row shows the basic performance of directly using all classification parameters for detection, without adaptation or knowledge transfer (*i.e.*, weakly supervised learning). The last row shows results of an oracle detection network which assumes that bounding boxes for all 200 categories are available (*i.e.*, supervised learning). The second row shows the baseline LSDA results using only feature adaptation. Rows 3-5 show the performance of LSDA for adapting both the feature layers (layer 1-7) and the class-specific layer (layer 8), by considering different numbers of neighbor categories. Rows 6-8, 9-12 and row 13 show the results of our visual transfer, semantic transfer and mixture transfer model, respectively. Row 14 shows our results after bounding-box regression. For all the methods, same "***AlexNet***" CNN is adopted.

Method	Number of Nearest Neighbors	mAP on $\mathcal{B}$: "Fully labeled" 100 Categories	mAP on $\mathcal{A}$: "Weakly labeled" 100 Categories	mAP on $\mathcal{D}$: All 200 Categories
Classification Network	-	12.63	10.31	11.90
LSDA (only class invariant adaptation)	-	27.81	15.85	21.83
LSDA (class invariant & specific adapt)	avg/weighted - 5	28.12 / –	15.97 / 16.12	22.05 / 22.12
	avg/weighted - 10	27.95 / –	16.15 / 16.28	22.05 / 22.12
	avg/weighted - 100	27.91 / –	15.96 / 16.33	21.94 / 22.12
Ours (visual transfer)	avg/weighted - 5	27.99 / –	17.42 / 17.59	22.71 / 22.79
	avg/weighted - 10	27.89 / –	17.62 / 18.41	22.76 / 23.15
	avg/weighted - 100	28.30 / –	17.38 / 19.02	22.84 / 23.66
Ours (semantic transfer)	avg/weighted - 5	28.01 / –	17.32 / 17.53	22.67 / 22.77
	avg/weighted - 10	28.00 / –	16.67 / 17.50	22.31 / 22.75
	avg/weighted - 100	28.14 / –	17.04 / 18.32	23.23 / 23.28
	Sparse rep. - $\leq$20	28.18	19.04	23.66
Ours (mixture transfer)	-	28.04	**20.03** ↑3.88	24.04
Ours (mixture transfer + BB reg.)	-	31.85	**21.88**	26.87
Oracle: Full Detection Network (no BB reg.)	-	29.72	26.25	28.00
Oracle: Full Detection Network (BB reg.)	-	32.17	29.46	30.82

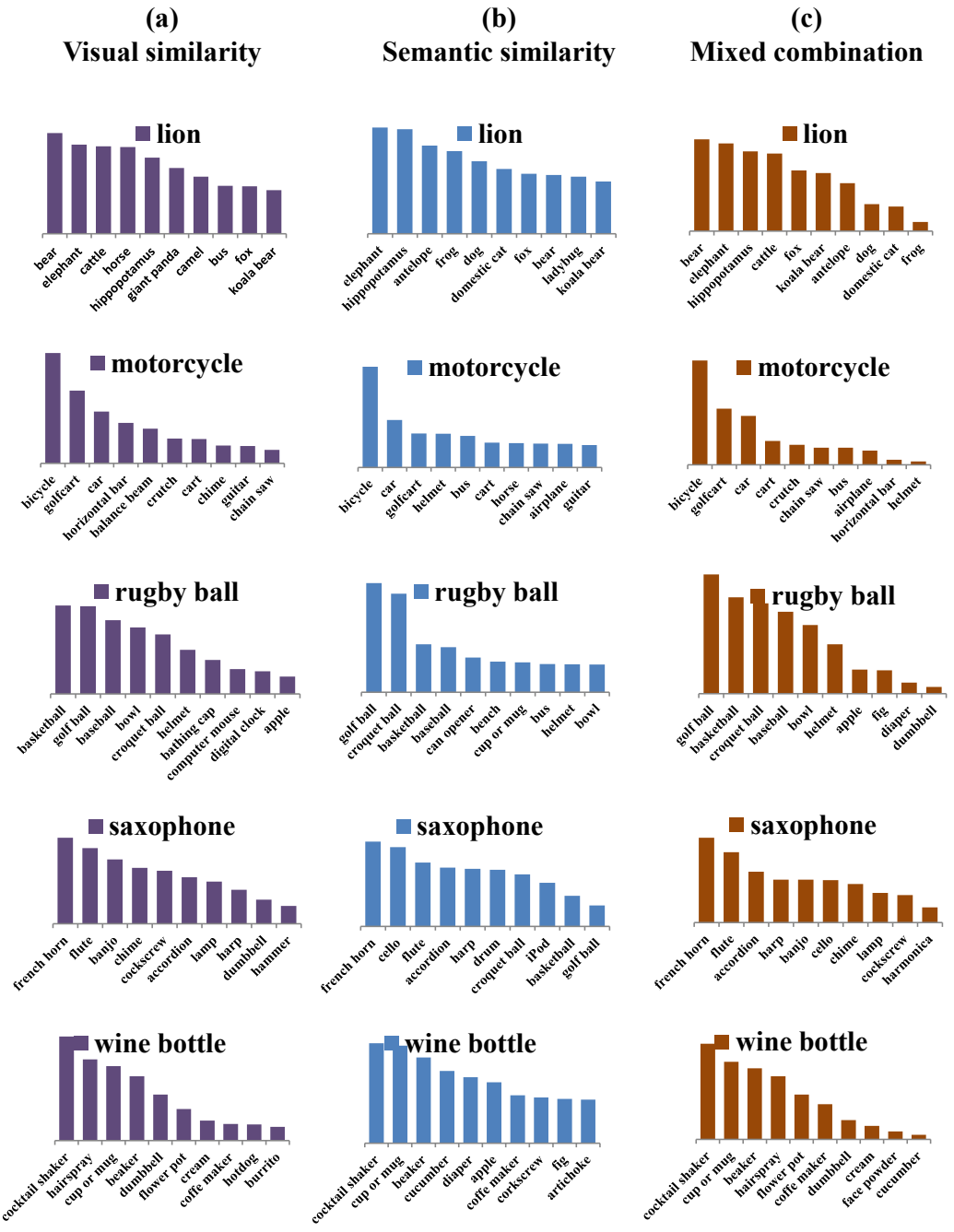

Figure 4.4: Some example visualizations of (a) visual similarity (first row in the figure), (b) semantic similarity (middle row) and (c) mixture similarity (last row) between a target "weakly labeled" category and its source categories from which to transfer knowledge. For each target category, the top-10 weighted nearest neighbor categories are shown. The magnitude of each column bar shows the relative weight (degree of similarity s_v, s_s, s in Eq. (4.6), where α is set to 0.6).

"*weighted*" scheme works steadily better than its "*average*" counterpart.

For our **visual knowledge transfer model**, we show steady improvement over the baseline LSDA methods when considering the average weight change of both 5 and 10 visually similar categories, with 1.45% and 1.47% increase in mAP, respectively. This proves that our proposed visual similarity measure is superior to that of LSDA, showing that category-specific adaptation can indeed be improved based on knowledge about the visual similarities between categories. Further improvement is achieved by modeling individual weights of all 100 source categories according to their degree of visual similarities to the target category (**weighted-100** with 19.02% in the table). This verifies our supposition that the transformation from a classifier to a detector of a certain category is more related to visually similar categories, and is proportional to their degrees of similarity. For example, *motorcycle* is most similar to *bicycle*. Thus the weight change from a *bicycle* classifier to detector has the largest influence on the transformation of *motorcycle*. The influence of less visually relevant categories, such as *cart* and *chain saw*, is much smaller. For visually dissimilar categories (*apple, fig, hotdog*, etc.), the influence is extremely insignificant. We show some examples of visual similarities between a target category and its source categories in the left column of Figure 4.4. For each target category, the top-10 weighted nearest neighbor categories with their similarity degrees are visualized.

Our **semantic knowledge transfer model** also showed marked improvement over the LSDA baseline (Table 4.1, Rows 9-12), and is comparable to the results of the visual transfer model. This suggests that the cross-domain knowledge transfer from semantic relatedness to visual similarity is very effective. The best performance for the semantic transfer model (19.04%) is obtained by sparsely reconstructing the target category with the source categories using the synset embeddings. The results of using synset embeddings (18.32%, using weighted-100, the same below) are superior to using 'naive' word2vec embeddings (17.83%) and WordNet based measures such as path-based similarity (17.08%) and Lin similarity [Lin 1998] (17.31%). Several examples visualizing the related categories of the 10 largest semantic reconstruction coefficients are shown in the middle column of Figure 4.4. We observe that semantic relatedness indeed correlates with visual sim-

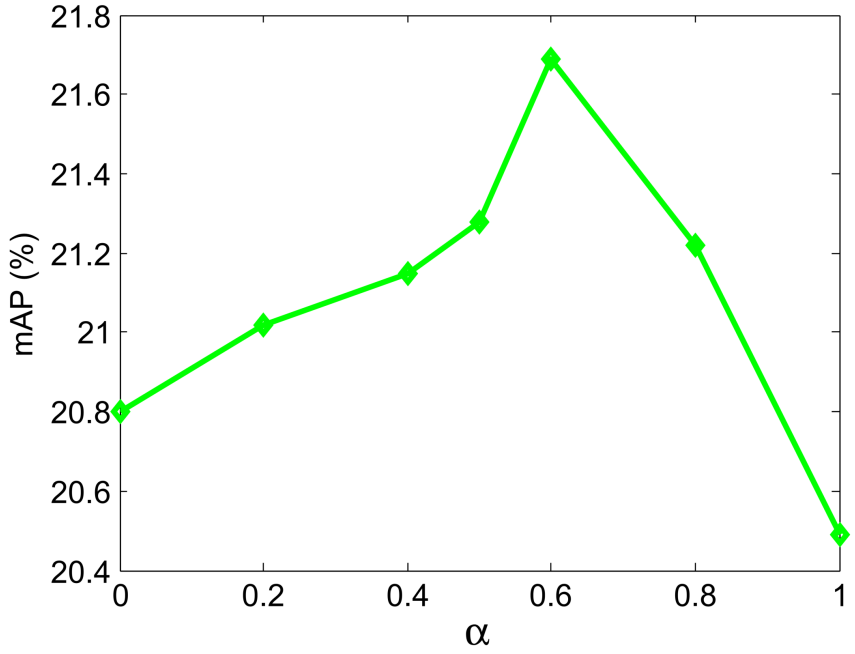

Figure 4.5: Sensitivity of parameter α vs. mAP. for detection of "weakly labeled" categories on the validation (val1) dataset. $\alpha \in [0, 1]$ is a parameter used to control the relative influence of the two similarity measurements. α is set to 1 when only considering visual similarity transfer, and 0 for the semantic similarity transfer.

ilarity.

Table 4.2: Comparison of mean average precision (mAP) for semantic similarity measures/representations, using **Weighted - 100**.

Method	Path Similarity	Lin Similarity	Naive Embeddings	AutoExtend (this paper)
mAP	17.08	17.31	17.83	18.32

The state-of-the-art result using the 8-layer "*Alex-Net*" for semi-supervised detection on this dataset is achieved by our **mixture transfer model** which combines visual similarity and semantic relatedness. A boost in performance of 3.88% on original split (3.82%±0.12%, based on 6 different splits of the dataset) is achieved over the best result reported by LSDA on the "weakly labeled" categories. We show examples of transferred categories with their corresponding weights for several target categories in the right column of Figure 4.4. The parameter α in

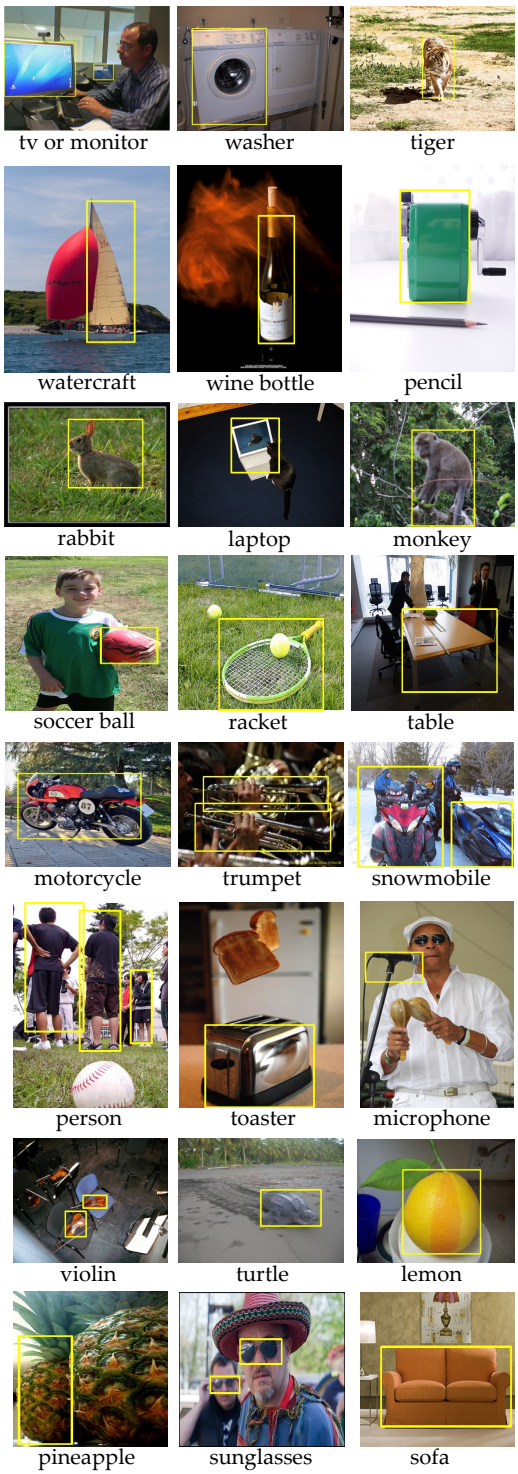

Figure 4.6: Examples of correct detections (true positives) of our mixture knowledge transfer model on ILSVRC2013 images. For each image, only detections for the "weakly labeled" target category (text below image) are listed.

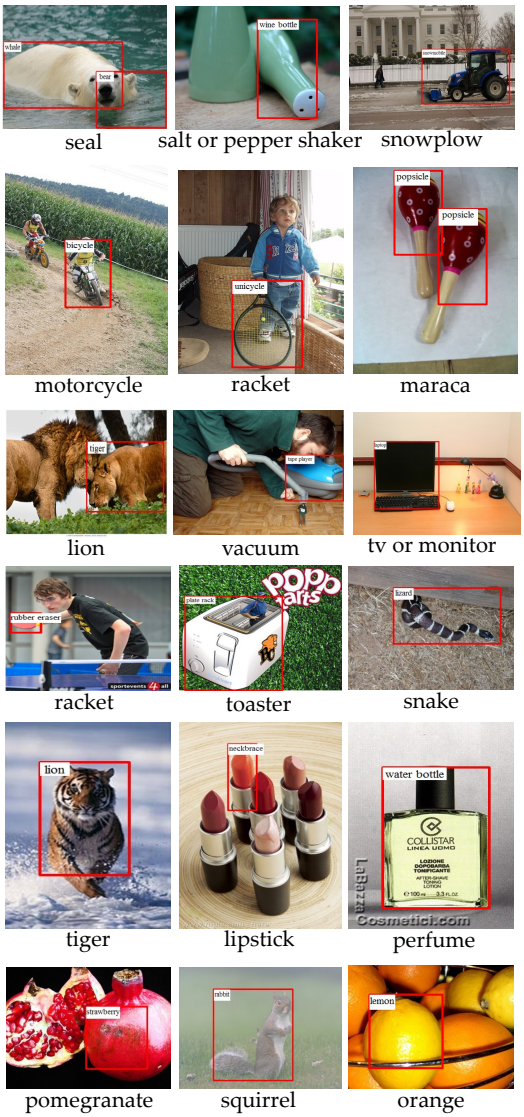

Figure 4.7: Examples of incorrect detections (confusion with other objects) of our mixture knowledge transfer model on ILSVRC2013 images. The detected object label is shown in the top-left of its bounding box.

Eq. (4.6) for the mixture model weights is set to 0.6 for final detection, where $\alpha \in \{0, 0.2, 0.4, 0.5, 0.6, 0.8, 1\}$ is chosen via cross-validation on the val1 detection set (Figure 4.5). This suggests that the transferring of visual similarities is slightly more important than semantic relatedness, although both are indeed complementary. We do not tune α for each category separately, though this can be expected to further improve our detection performance. Figures 4.6 and 4.7 show some examples of correct and incorrect detections respectively. Although our proposed mixture transfer model achieves the state-of-the-art in detecting the "weakly labeled" categories, it is still occasionally confused by visually similar categories.

4.4.4 Experimental Results with "*VGG-Nets*"

Previous work [Simonyan & Zisserman 2015, Girshick *et al.* 2016, He *et al.* 2016] found that region based CNN detection performance is significantly influenced by the choice of CNN architecture. In Table 4.3, we show some detection results using the 16-layer and 19-layer deep "*VGG-Nets*" proposed by Simonyan and Zisserman [Simonyan & Zisserman 2015]. The *VGG-16* network is consisted of 13 convolutional layers of very small (3×3) convolution filters, with 5 max pooling layers interspersed, and topped with 3 fully co nnected layers (namely, $fc6$, $fc7$ and $fc8$). The *VGG-19* network extends *VGG-16* by inserting 3 more convolutional layers, while keeping other layer configurations unchanged.

As can be seen from Table 4.3, the very deep ConvNets *VGG-16* and *VGG-19* significantly outperform the *Alex-Net* for all the adaptation methods. Our knowledge transfer models using the very deep *VGG-nets* with different similarity measures show consistent improvement over the LSDA basedline methods. The overall improvement over performance using the *VGG-Net* is similar with that of the *Alex-Net*. In principle, We would expect further improvement by using the *Res-Net* [He *et al.* 2016] which has more than 150 layers. However, testing various deeper networks is out of scope for this paper.

Table 4.3: Comparison of detection mean average precision (mAP) on the "weakly labeled" categories of ILSVRC2013 val2, using the "***VGG-Nets***". For LSDA, our visual similarity and semantic relatedness transfer models, **Weighted - 100** scheme is adopted.

Method	Only classification	LSDA class inv.	LSDA class inv. & spec.	
Alex-Net	10.31	15.85	16.33	
VGG-16	14.89	18.24	18.86	
VGG-19	16.22	20.38	21.02	
Method	**Ours visual**	**Ours semantic**	**Ours mixed**	**Ours mixed + BB reg.**
Alex-Net	19.02	18.32	20.03	21.88
VGG-16	21.75	21.07	23.21	24.91
VGG-19	23.89	23.10	25.07	27.32

4.4.5 Experimental Results with Bounding-box Regression

Results in Table 4.1 and Table 4.3 show that the transferred bounding-box regression from "fully labeled" categories fixes a large number of detections resulted by mis-localization, boosting mAP by about 2 points for the "weakly labeled" categories. The bounding-box regression process could boost mAP by 3 to 4 points if the bounding box annotations for all the categories were provided. We show some example detections before and after bounding box regression on the "weakly labeled" categories in Fig. 4.8, using *VGG-16*.

4.4.6 Detection Error Analysis

We present an analysis of the types of errors that our models make. We use the detection diagnosis tool of [Hoiem *et al.* 2012] and consider three types of false positive (FP) errors: *Loc* (poor localizations), *Oth* (confusion with other objects, *e.g.*, correctly localizing an object but classifying it to a wrong class) and *BG* (confusion with background or unlabeled objects). We discard the *Sim* (confusion with similar objects) errors, since the ground-truth similarities between objects categories

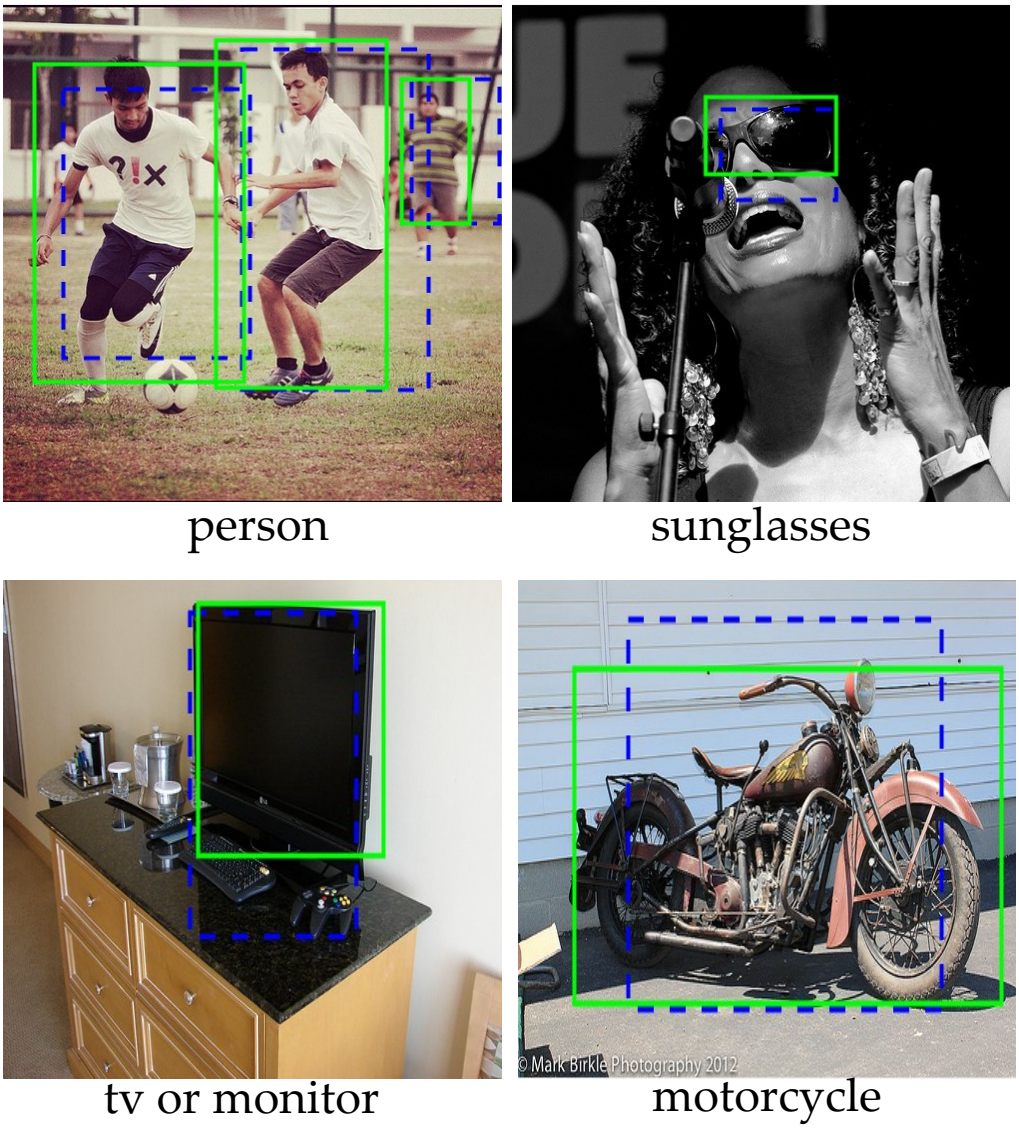

Figure 4.8: Some example detections before and after bounding box regression on the "weakly labeled" categories. Boxes before (resp. after) bounding-box regression are shown in dashed blue (resp. green).

are not explicitly defined in ILSVRC dataset.

We show the distribution of top-ranked (top scoring 25 to 3200) false positive (FP) types for our mixture knowledge transfer model with *VGG-16* on the weakly labeled categories in Figure. 4.9. s can be seen from Figure. 4.9 (a) and (b), a majority of our errors result from confusion with other object categories, indicating that the knowledge transfer model might be influenced by other categories. We noticed

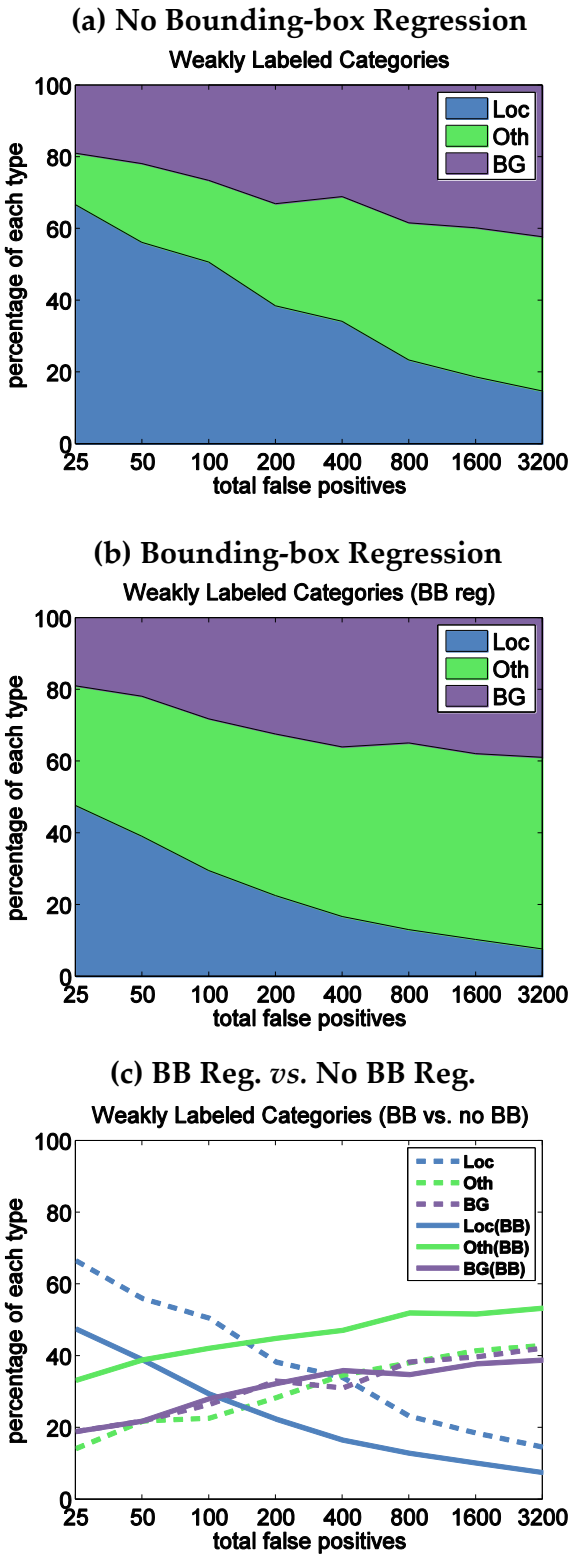

Figure 4.9: Analysis of detection errors of our model. Error trend and fractions before and after bounding-box regression are compared. Best viewed in color.

that many of *Oth* errors were from the categories similar to the target category (See Figure. 4.7). We can see the blue curves from Figure. 4.9 (c) that the bounding-box regression transfer method is very effective at fixing localization errors.

4.5 Summary

In this chapter, we investigated how knowledge about object similarities from both visual and semantic domains can be transferred to adapt an image classifier to an object detector in a semi-supervised setting. We experimented with different CNN architectures on the challenging ILSVRC2013 detection dataset, found clear evidence that both visual and semantic similarities play an essential role in improving the adaptation process, and that the combination of the two modalities yielded state-of-the-art performance, suggesting that knowledge inherent in visual and semantic domains is complementary.

Chapter 5

Conclusion and Future Work

Contents

5.1 Summary of Contributions 109
5.2 Perspective for Future Directions 111

While fully supervised learning methods achieve best detection performance in general, they rely too heavily on large-scale datasets with careful object-level annotations. Although object-level annotations are extremely valuable, the process of manually annotating object bounding boxes on large-scale dataset is often time consuming, expensive, and not-trivial to setup. In this dissertation we have focused on the problem of weakly supervised object detection, where the object-level annotations are incomplete in the training stage. We proposed two approaches for addressing such cases: (i) We presented a region proposal-selection framework, building on the Deformable Part-based Models (DPMs) for weakly supervised object detection, given only image-level labels in training. (ii) We investigated how knowledge about object similarities from both visual and semantic domains can be transferred to adapt an image classifier to an object detector in a semi-supervised setting, where only a subset of object categories are annotated with bounding boxes. The following concludes the thesis with a summary of contributions in Section 5.1, and potential directions for future research in Section 5.2.

5.1 Summary of Contributions

In Chapter 3, we proposed a model enhancing weakly supervised learning by emphasizing the importance of location and size of the initial class specific root

filter of deformable part-based models (DPMs). We follow the general setup of [Pandey & Lazebnik 2011] and introduce several substantial improvements to the weakly supervised deformable part-based model (DPMs). The main contributions included a new selection model based on generic "objectness" (region proposals) and visual saliency to adaptively select a reliable set of candidate windows which tend to represent the object instances in the image, and a latent class learning process by coarsely classifying a candidate window into either a target object or a non-target class using image-level CNN classifiers which was pre-trained on the large-scale ImageNet dataset and fine-tuned for on PASCAL VOC for classification. Furthermore, we designed a flexible enlarging-and-shrinking post-processing procedure to modify the output bounding boxes of DPMs, which can effectively generate more accurate bounding boxes by better conserving foreground and cropping out plain background regions, which aims to approximatively match the object aspect ratios, to further improve the final accuracy. Moreover, we incorporate the contextual information from image classification, by combining the image-level classification score with object-level DPM detection score, to obtain a final score for detection. The proposed multiple region initialization method can detect multiple co-existing objects in the image. Experimental results on multiple datasets demonstrate that our proposed framework is efficient and competitive with the state-of-the-art, especially for the object categories which are relatively salient and deformable. The proposed initialization method for selecting candidate windows can be also utilized by other kinds of detectors for weakly supervised object detection.

In Chapter 4, we investigated how knowledge about object similarities from both visual and semantic domains can be transferred to adapt an image classifier to an object detector in a semi-supervised setting, where only a subset of object categories are annotated with bounding boxes. We defined a visual similarity measurement based on visual appearance and a semantic similarity measurement based on Word2vec embeddings. We modeled the category-specific differences between CNN classifiers and CNN detectors on "fully labeled" categories, by considering the degree of similarities between (visual or semantic) feature vectors, and subse-

quently transferring this knowledge for adapting an image classifier to an object detector for a "weakly labeled" category. We experimented with different CNN architectures, found clear evidence that both visual and semantic similarities play an essential role in improving the adaptation process, and that the combination of the two modalities yielded state-of-the-art performance, suggesting that knowledge inherent in visual and semantic domains is complementary. We also found these knowledge can be transferred to the bounding box regression process for weakly labeled categories to achieve better performance. The main advantage of our proposed method is that knowledge about object similarities can be obtained without requiring further object-level annotations, for example from existing image databases, text corpora and external knowledge bases.

In general, these contributions of this dissertation bring us significantly closer to the goal of scalable learning of strong models from weakly annotated non-purpose collected data on the Internet.

5.2 Perspective for Future Directions

Based on the work presented in this dissertation, we present several possible research directions.

1. *Effective and efficient detection of small objects in a weakly supervised manner.* Small objects can be appeared in the real-world images and they are very hard to detect even when object level annotations are given in the training process. The proposed multiple-region initialization DPMs model has the ability to detect multiple objects in the images, however, its performance on small objects is not desirable. Similar for our CNN-based approach, effective and efficient detection of small objects is one of the most interesting bit of (weakly supervised) object detection. A possible method is to obtain a more powerful feature map representation of an image or an image region by using smaller convolutional filters or smaller stride, however, this may probably raise the problem of inefficiency in training such an expensive network. Therefore further research is needed in order to solve this problem in an efficient way.

2. *End-to-end learning of Convolutional Neural Networks for weakly supervised object detection.* A number of methods [Ren *et al.* 2015a, Redmon *et al.* 2016, Liu *et al.* 2016] have been proposed to learn a CNN detector in an end-to-end structure to accelerate the detection speed in a fully supervised manner. However, this remains challenging for weakly supervised training from image-level labels. For the existed weakly supervised CNN-based methods [Wang *et al.* 2015, Bilen & Vedaldi 2016], external computation of region proposals (*e.g.* , Selective Search [Uijlings *et al.* 2013]) are needed. Yet when compared to efficient detection networks, Selective Search is an order of magnitude slower, at 2 seconds per image in a CPU implementation. EdgeBoxes [Zitnick & Dollar 2014] region proposal currently provides the best tradeoff between proposal quality and speed, at 0.2 seconds per image. However, the region proposal step still consumes as much running time as the detection network. Therefore, an end-to-end learning framework is desired for weakly supervised object detection, which can generate accurate region proposals based on the feature maps automatically, to accelerate the training and detection speed, along with improved performance.

3. *Learning from noisy data.* It is desirable that weakly supervised object detectors could deal with noisy images or annotations. This dissertation did not provide a thorough analysis on learning from noisy data. However, in the real-world, the collected image-level labels can be even noisy due to annotator bias (*e.g.* , missing objects, confusion with other categories). It is desirable to investigate the robust capacity of the proposed model or related extended models.

4. *Exploration of transferable knowledge from different contents or domains.* More domains or contents (*e.g.* , video, sound, text description of the visual content, attribute) using better representations can be investigated to help weakly supervised learning for detection. We believe that the combination of knowledge from different domains is key to improving weakly supervised or semi-supervised object detection. In addition, object detectors should be able to extend to previously unseen categories based on the transferable knowledge. It is also very important to investigate contextual information from various domains to help object detection. Moreover, one of the future research directions is to investigate the possibil-

ity of using category-invariant properties, for example, by modeling the difference between feature distributions of whole images and target objects, to transform a classification network to a detection network using back propagation.

We can only see a short distance ahead, but we can see plenty there that needs to be done.

-Alan Turing, *Computing Machinery and Intelligence (1950)*

Chapter 6

List of Publications

Articles in peer-reviewed journals

- Yuxing Tang, Xiaofang Wang, Emmanuel Dellandréa, Liming Chen. *Weakly Supervised Learning of Deformable Part-Based Models for Object Detection via Region Proposals.* IEEE Transactions on Multimedia (TMM), accepted, to appear in January 2017.

- Xiaofang Wang, Yuxing Tang, Simon Masnou, Liming Chen. *A Global/Local Affinity Graph for Image Segmentation.* IEEE Transactions on Image Processing (TIP), vol. 24, no. 4, pages 1399–1411, 2015.

International peer-reviewed conferences

- Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, 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 (CVPR), Las Vegas, 2016.

- Yuxing Tang, Xiaofang Wang, Emmanuel Dellandréa, Simon Masnou, Liming Chen. *Fusing Generic Objectness and Deformable Part-based Models for Weakly Supervised Object Detection.* In Proceedings of the IEEE International Conference on Image Processing (ICIP), Paris, 2014. (Oral, top 10% award)

- Yuxing Tang, Charles-Edmond, Chao Zhu. *Fan-shaped Patch Local Binary Patterns for Texture Classification.* In Proceedings of the 11th International Workshop on Content-Based Multimedia Indexing (CBMI), Veszprem, 2013. (Oral)

Other publications

- Hervé Le Borgne, ..., Yuxing Tang, Emmanuel Dellandréa, Charles-Edmond Bichot, Liming Chen. *IRIM at TRECVID 2015: Semantic Indexing.* In Proceedings of the TREC Video Retrieval Evaluation 2015 workshop, Gaithersburg, MD, United States, 2015.

- Nicolas Ballas, ..., Yuxing Tang, Emmanuel Dellandréa, Charles-Edmond Bichot, Liming Chen. *IRIM at TRECVID 2014: Semantic Indexing and Instance Search.* In Proceedings of the TREC Video Retrieval Evaluation 2014 workshop, Orlando, FL, United States, 2014.

- Nicolas Ballas, ..., Yuxing Tang, Emmanuel Dellandréa, Charles-Edmond Bichot, Liming Chen. *IRIM at TRECVID 2013: Semantic Indexing and Instance Search.* In Proceedings of the TREC Video Retrieval Evaluation 2013 workshop, Gaithersburg, MD, United States, 2013.

- Nicolas Ballas, ..., Yuxing Tang, Emmanuel Dellandréa, Charles-Edmond Bichot, Liming Chen. *IRIM at TRECVID 2013: Semantic Indexing and Instance Search.* Proceedings of TREC Video Retrieval Evaluation 2012 workshop, Gaithersburg, MD, United States, 2012.

Article under preparation

- Yuxing Tang, Josiah Wang, Xiaofang Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, Liming Chen. *Large Scale Semi-supervised Object Detection Using Visual and Semantic Knowledge Transfer.* In preparation for submission to IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

Bibliography

[Agrawal *et al.* 2014] P. Agrawal, R. Girshick and J. Malik. *Analyzing the Performance of Multilayer Neural Networks for Object Recognition*. In Proceedings of the European Conference on Computer Vision (ECCV), 2014. 90

[Ahonen *et al.* 2006] T. Ahonen, A. Hadid and M. Pietikainen. *Face Description with Local Binary Patterns: Application to Face Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 28, no. 12, pages 2037–2041, 2006. 16

[Alexe *et al.* 2012] B. Alexe, T. Deselaers and V. Ferrari. *Measuring the Objectness of Image Windows*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 34, no. 11, pages 2189–2202, 2012. 6, 8, 29, 37, 38, 47, 49, 50

[Amit & Felzenszwalb 2014] Y. Amit and P. Felzenszwalb. *Object Detection*. In Computer Vision: A Reference Guide. Springer, 2014. 20

[Andreopoulos & Tsotsos 2013] A. Andreopoulos and J. K. Tsotsos. *50 Years of Object Recognition: Directions Forward*. Computer Vision and Image Understanding (CVIU), vol. 117, no. 8, pages 827–891, 2013. 3, 14

[Andrews *et al.* 2003] S. Andrews, I. Tsochantaridis and T. Hofmann. *Support Vector Machines for Multiple-Instance Learning*. In Proceedings of Advances in Neural Information Processing Systems (NIPS). 2003. 36, 71

[Arbeláez *et al.* 2014] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques and J. Malik. *Multiscale Combinatorial Grouping*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 29, 50

[Attardi *et al.* 2009] G. Attardi, F. Dell'Orletta, M. Simi and J. Turian. *Accurate Dependency Parsing with a Stacked Multilayer Perceptron*. In Proceeding of Evalita, 2009. 19

[Azizpour & Laptev 2012] H. Azizpour and I. Laptev. *Object Detection Using Strongly-Supervised Deformable Part Models*. In Proceedings of the European Conference on Computer Vision (ECCV), 2012. 46

[Bay *et al.* 2008] H. Bay, A. Ess, T. Tuytelaars and L. Van Gool. *Speeded-up robust features (SURF)*. Computer Vision and Image Understanding (CVIU), vol. 110, no. 3, pages 346–359, 2008. 17

[Belongie *et al.* 2002] S. Belongie, J. Malik and J. Puzicha. *Shape Matching and Object Recognition Using Shape Contexts*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 24, no. 4, pages 509–522, 2002. 17

[Bengio *et al.* 2007] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle. *Greedy Layer-wise Training of Deep Networks*. In Proceedings of Advances in Neural Information Processing Systems (NIPS), 2007. 40

[Bilen & Vedaldi 2016] H. Bilen and A. Vedaldi. *Weakly Supervised Deep Detection Networks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 39, 112

[Bilen *et al.* 2014] H. Bilen, M. Pedersoli and T. Tuytelaars. *Weakly Supervised Object Detection with Posterior Regularization*. In Proceedings of the British Machine Vision Conference (BMVC), 2014. 8, 73

[Bilen *et al.* 2015] H. Bilen, M. Pedersoli and T. Tuytelaars. *Weakly Supervised Object Detection With Convex Clustering*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 38, 69, 71, 73

[Bishop 1995] C. M. Bishop. Neural networks for pattern recognition. Oxford University Press, Inc., New York, NY, USA, 1995. 19

[Borji *et al.* 2015] A. Borji, M. Cheng, H. Jiang and J. Li. *Salient Object Detection: A Benchmark*. IEEE Transactions on Image Processing (TIP), vol. 24, no. 12, pages 5706–5722, 2015. 37

[Boureau *et al.* 2010] Y. L. Boureau, J. Ponce and Y. LeCun. *A Theoretical Analysis of Feature Pooling in Vision Algorithms*. In Proceedings of the International Conference on Machine learning (ICML), 2010. 18

[Carreira & Sminchisescu 2012] J. Carreira and C. Sminchisescu. *CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 34, no. 7, pages 1312–1328, July 2012. 29, 50

[Carreira *et al.* 2012] J. Carreira, R. Caseiro, J. Batista and C. Sminchisescu. *Semantic Segmentation with Second-order Pooling*. In Proceedings of the European Conference on Computer Vision (ECCV), 2012. 18

[Chang & Lin 2011] C. C. Chang and C. J. Lin. *LIBSVM: A Library for Support Vector Machines*. ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, pages 27:1–27:27, May 2011. 56

[Chatfield *et al.* 2011] K. Chatfield, V. Lempitsky, A. Vedaldi and A. Zisserman. *The Devil is in the Details: An Evaluation of Recent Feature Encoding Methods*. In Proceedings of the British Machine Vision Conference (BMVC), 2011. 18

[Cheng *et al.* 2014] M. M. Cheng, Z. Zhang, W. Y. Lin and P. Torr. *BING: Binarized Normed Gradients for Objectness Estimation at 300fps*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 29, 50

[Cho *et al.* 2015] M. Cho, S. Kwak, C. Schmid and J. Ponce. *Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 40

[Cinbis *et al.* 2014] R.G. Cinbis, J. Verbeek and C. Schmid. *Multi-fold MIL Training for Weakly Supervised Object Localization*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 8, 38, 39, 72, 73

[Cortes & Vapnik 1995] C. Cortes and V. Vapnik. *Support-vector Networks*. Machine Learning (ML), vol. 20, no. 3, pages 273–297, 1995. 20, 21

[Csurka *et al.* 2004] G. Csurka, C. Dance, L. Fan, J. Willamowski and C. Bray. *Visual Categorization with Bags of Keypoints*. In Workshop on statistical learning in computer vision, Proceedings of the European Conference on Computer Vision (ECCVW), 2004. 18, 19, 72

[Dalal & Triggs 2005] N. Dalal and B. Triggs. *Histograms of Oriented Gradients for Human Detection*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005. 4, 5, 6, 17, 24, 25, 45, 46, 57

[Daugman 1988] J. G. Daugman. *Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression*. IEEE Transactions on Acoustics, Speech, and Signal Processing (TASPSP), vol. 36, no. 7, pages 1169–1179, 1988. 15

[Deselaers & Ferrari 2011] T. Deselaers and V. Ferrari. *Visual and Semantic Similarity in ImageNet*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011. 85, 91

[Deselaers *et al.* 2012] T. Deselaers, B. Alexe and V. Ferrari. *Weakly Supervised Localization and Learning with Generic Knowledge*. International Journal of Computer Vision (IJCV), vol. 100, no. 3, pages 275–293, 2012. 38, 62, 63, 64, 65, 66, 69, 71

[Donahue *et al.* 2013] J. Donahue, J. Hoffman, E. Rodner, K. Saenko and T. Darrell. *Semi-Supervised Domain Adaptation with Instance Constraints*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013. 39

[Dubout & Fleuret 2012] C. Dubout and F. Fleuret. *Exact Acceleration of Linear Object Detectors*. In Proceedings of the European Conference on Computer Vision (ECCV), 2012. 28

[Everingham *et al.* 2010] 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 (IJCV), vol. 88, no. 2, pages 303–338, 2010. 4, 6, 7, 8, 26, 28, 46, 48, 62, 83

[Fellbaum 1998] C. Fellbaum. Wordnet: An electronic lexical database. MIT Press, Cambridge, MA, 1998. 91

[Felzenszwalb *et al.* 2010a] P. Felzenszwalb, R. Girshick and D. McAllester. *Cascade Object Detection with Deformable Part Models*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010. 28

[Felzenszwalb *et al.* 2010b] P. F. Felzenszwalb, R. Girshick, D. McAllester and D. Ramanan. *Object Detection with Discriminatively Trained Part Based Models*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 32, no. 9, pages 1627–1645, 2010. 5, 6, 9, 24, 25, 27, 31, 45, 46, 57, 59, 73, 74, 76

[Fergus *et al.* 2003] R. Fergus, P. Perona and A. Zisserman. *Object Class Recognition by Unsupervised Scale-invariant Learning*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2003. 21

[Fernando *et al.* 2012] B. Fernando, E. Fromont, D. Muselet and M. Sebban. *Supervised Learning of Gaussian Mixture Models forVisual Vocabulary Generation*. Pattern Recognition (PR), vol. 45, no. 2, pages 897–907, 2012. 18

[Foresti *et al.* 2002] G. L. Foresti, L. Marcenaro and C. S. Regazzoni. *Automatic Detection and Indexing of Video-event Shots for Surveillance Applications*. IEEE Transactions on Multimedia (TMM), vol. 4, no. 4, pages 459–471, 2002. 45

[Freund & Schapire 1997] Y. Freund and R. E. Schapire. *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*. Journal of Computer and System Sciences (JCSS), vol. 55, no. 1, pages 119–139, 1997. 21

[Galleguillos *et al.* 2008] C. Galleguillos, B. Babenko, A. Rabinovich and S. Belongie. *Weakly Supervised Object Recognition and Localization with Stable Seg-*

mentations. In Proceedings of the European Conference on Computer Vision (ECCV), 2008. 36

[Gao *et al.* 2012] B. Gao, E. Dellandrea and L. Chen. *Music Sparse Decomposition onto a MIDI Dictionary of Musical Words and Its Application to Music Mood Classification*. In International Workshop on Content-Based Multimedia Indexing (CBMI), 2012. 93

[Girshick *et al.* 2011] R. Girshick, P. F. Felzenszwalb and D. McAllester. *Object Detection with Grammar Models*. In Proceedings of Advances in Neural Information Processing Systems (NIPS), 2011. 46

[Girshick *et al.* 2014] R. Girshick, J. Donahue, T. Darrell and J. Malik. *Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 5, 6, 22, 23, 24, 30, 31, 45, 50, 55, 58, 73, 74, 82, 84, 87, 94, 96

[Girshick *et al.* 2015] R. Girshick, F. Iandola, T. Darrell and J. Malik. *Deformable Part Models are Convolutional Neural Networks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 57, 58, 68, 73, 74

[Girshick *et al.* 2016] R. Girshick, J. Donahue, T. Darrell and J. Malik. *Region-Based Convolutional Networks for Accurate Object Detection and Segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 38, no. 1, pages 142–158, 2016. 9, 22, 30, 104

[Girshick 2015] R. Girshick. *Fast R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. In Proceedings of the International Conference on Computer Vision (ICCV), 2015. 6, 9, 22, 30, 33, 34, 82

[Harris & Stephens 1988] C. Harris and M. Stephens. *A Combined Corner and Edge Detector*. In Proceedings of The Alvey Vision Conference, 1988. 16

[He *et al.* 2015] K. He, X. Zhang, S. Ren and J. Sun. *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 37, no. 9, pages 1904–1916, 2015. 30, 31, 32, 58, 82

[He *et al.* 2016] K. He, X. Zhang, S. Ren and J. Sun. *Deep Residual Learning for Image Recognition*. In Proceedings of the The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 22, 23, 104

[Hoffman *et al.* 2014] J. Hoffman, S. Guadarrama, E. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell and K. Saenko. *LSDA: Large Scale Detection through Adaptation*. In Proceedings of Advances in Neural Information Processing Systems (NIPS), 2014. 39, 40, 83, 84, 86, 88, 89, 96

[Hoiem *et al.* 2012] D. Hoiem, Y. Chodpathumwan and Q. Dai. *Diagnosing Error in Object Detectors*. In Proceedings of the European Conference on Computer Vision (ECCV), 2012. 74, 75, 105

[Hosang *et al.* 2014] J. Hosang, R. Benenson and B. Schiele. *How Good Are Detection Proposals, Really?* In Proceedings of the British Machine Vision Conference (BMVC), 2014. 28, 50, 65

[Hosang *et al.* 2016] J. Hosang, R. Benenson, P. Dollar and B. Schiele. *What Makes for Effective Detection Proposals?* IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 38, no. 4, pages 814–830, 2016. 30

[Hyvärinen & Oja 2000] A. Hyvärinen and E. Oja. *Independent Component Analysis: Algorithms and Applications*. Neural Networks (NN), vol. 13, no. 4–5, 2000. 19

[Jia *et al.* 2014] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama and T. Darrell. *Caffe: Convolutional Architecture for Fast Feature Embedding*. arXiv preprint arXiv:1408.5093, 2014. 55, 90, 96

[Kadir & Brady 2001] T. Kadir and M. Brady. *Saliency, Scale and Image Description*. International Journal of Computer Vision (IJCV), vol. 45, no. 2, pages 83–105, 2001. 16

[Ke *et al.* 2006] Y. Ke, X. Tang and F. Jing. *The Design of High-Level Features for Photo Quality Assessment*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006. 61

[Kim & Torralba 2009] G. Kim and A. Torralba. *Unsupervised Detection of Regions of Interest using Iterative Link Analysis*. In Proceedings of Advances in Neural Information Processing Systems (NIPS), 2009. 37

[Kokkinos 2011] I. Kokkinos. *Rapid Deformable Object Detection using Dual-Tree Branch-and-Bound*. In Proceedings of Advances in Neural Information Processing Systems (NIPS). 2011. 28

[Krizhevsky *et al.* 2012] A. Krizhevsky, I. Sutskever and G. E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks*. In Proceedings of Advances in Neural Information Processing Systems (NIPS). 2012. 6, 14, 19, 22, 23, 24, 55, 82, 87

[Lafferty *et al.* 2001] J. D. Lafferty, A. McCallum and F. C. N. Pereira. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. In Proceedings of the International Conference of Machine Learning (ICML), 2001. 39

[Lazebnik *et al.* 2006] S. Lazebnik, C. Schmid and J. Ponce. *Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006. 18

[LeCun *et al.* 1990] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel. *Handwritten Digit Recognition with A Back-propagation Network*. In Proceedings of Advances in Neural Information Processing Systems (NIPS), 1990. 19, 22

[Lecun *et al.* 1998] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner. *Gradient-based Learning Applied to Document Recognition*. Proceedings of the IEEE, vol. 86, no. 11, pages 2278–2324, 1998. 19, 22

[Lee *et al.* 2006] H. Lee, A. Battle, R. Raina and A. Y. Ng. *Efficient Sparse Coding Algorithms*. In Proceedings of Advances in Neural Information Processing Systems (NIPS), 2006. 19

[Li *et al.* 2013] H. Li, F. Meng and K. N. Ngan. *Co-Salient Object Detection From Multiple Images*. IEEE Transactions on Multimedia (TMM), vol. 15, no. 8, pages 1896–1909, 2013. 51

[Li *et al.* 2016] Y. Li, L. Liu, C. Shen and A. van den Hengel. *Image Co-localization by Mimicking A Good Detector's Confidence Score Distribution*. In Proceedings of the European Conference on Computer Vision (ECCV), 2016. 40

[Lin *et al.* 2014a] M. Lin, Q. Chen and S. Yan. *Network In Network*. In International Conference on Learning Representations (ICLR), 2014. 23

[Lin *et al.* 2014b] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar and L. Zitnick. *Microsoft COCO: Common Objects in Context*. In Proceedings of the European Conference on Computer Vision (ECCV), 2014. 2, 4, 8, 23, 28, 48, 62, 78, 83

[Lin 1998] D. Lin. *An Information-Theoretic Definition of Similarity*. In Proceedings of the International Conference of Machine Learning (ICML), 1998. 100

[Lindeberg 1998] T. Lindeberg. *Feature Detection with Automatic Scale Selection*. International Journal of Computer Vision (IJCV), vol. 30, no. 2, pages 79–116, 1998. 16

[Liu *et al.* 2016] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu and A. Berg. *SSD: Single Shot MultiBox Detector*. In Proceedings of the European Conference on Computer Vision (ECCV), 2016. 6, 24, 29, 30, 82, 112

[Lowe 1999] D. G. Lowe. *Object Recognition from Local Scale-Invariant Features*. In Proceedings of the International Conference on Computer Vision (ICCV), 1999. 72

[Lowe 2001] D. G. Lowe. *Local Feature View Clustering for 3D Object Recognition*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2001. 16

[Lowe 2004] D. G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. International Journal of Computer Vision (IJCV), vol. 60, no. 2, pages 91–110, 2004. 16, 17

[Macqueen 1967] J. Macqueen. *Some Methods for Classification and Analysis of Multivariate Observations*. In Proceedings of Berkeley Symposium on Mathematical Statistics and Probability, 1967. 18

[Mairal *et al.* 2009] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman and Francis R. B. *Supervised Dictionary Learning*. In Proceedings of Advances in Neural Information Processing Systems (NIPS). 2009. 19

[Makhzani & Frey 2014] A. Makhzani and B. J. Frey. *k-Sparse Autoencoders*. In International Conference on Learning Representations (ICLR), 2014. 19

[Maron & Ratan 1998] O. Maron and A. L. Ratan. *Multiple-Instance Learning for Natural Scene Classification*. In Proceedings of the International Conference of Machine Learning (ICML), 1998. 8, 36

[Matas *et al.* 2004] J. Matas, O. Chum, M. Urban and T. Pajdla. *Robust wide-baseline stereo from maximally stable extremal regions*. Image and vision computing (IVC), vol. 22, no. 10, pages 761–767, 2004. 16

[Mikolajczyk & Schmid 2002] K. Mikolajczyk and C. Schmid. *An Affine Invariant Interest Point Detector*. In Proceedings of the European Conference on Computer Vision (ECCV), 2002. 16

[Mikolajczyk & Schmid 2004] K. Mikolajczyk and C. Schmid. *Scale & Affine Invariant Interest Point Detectors*. International Journal of Computer Vision (IJCV), vol. 60, no. 1, pages 63–86, 2004. 16

[Mikolajczyk *et al.* 2004] K. Mikolajczyk, C. Schmid and A. Zisserman. *Human Detection Based on a Probabilistic Assembly of Robust Part Detectors*. In Proceedings of the European Conference on Computer Vision (ECCV), 2004. 16

[Mikolajczyk *et al.* 2005] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool. *A Comparison of Affine Region Detectors*. International Journal of Computer Vision (IJCV), vol. 65, no. 1-2, pages 43–72, 2005. 17

[Mikolov *et al.* 2013a] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean. *Distributed Representations of Words and Phrases and their Compositionality*. In Proceedings of Advances in Neural Information Processing Systems (NIPS), 2013. 91, 97

[Mikolov *et al.* 2013b] Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig. *Linguistic Regularities in Continuous Space Word Representations*. In Proceedings of The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2013. 91

[Misra *et al.* 2015] I. Misra, A. Shrivastava and M. Hebert. *Watch and Learn: Semi-Supervised Learning of Object Detectors from Videos*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 86

[Murase & Nayar 1995] H. Murase and S. K. Nayar. *Visual Learning and Recognition of 3-D Objects from Appearance*. International Journal of Computer Vision (IJCV), vol. 14, no. 1, pages 5–24, 1995. 15

[Nair & Hinton 2010] V. Nair and G. E. Hinton. *Rectified Linear Units Improve Restricted Boltzmann Machines*. In Proceedings of the International Conference on Machine Learning (ICML), pages 807–814, 2010. 22

[Nascimento & Marques 2006] J. C. Nascimento and J. S. Marques. *Performance Evaluation of Object Detection Algorithms for Video Surveillance*. IEEE Transactions on Multimedia (TMM), vol. 8, no. 4, pages 761–774, 2006. 45

[Nguyen *et al.* 2009] M. Nguyen, L. Torresani, F. de la Torre and C. Rother. *Weakly Supervised Discriminative Localization and Classification: A Joint Learning Process*. In Proceedings of the International Conference on Computer Vision (ICCV), 2009. 36, 69, 71

[Nilsson 2009] N. J. Nilsson. The Quest for Artificial Intelligence. Cambridge University Press, New York, NY, USA, 2009. 2

[Nowak *et al.* 2006] E. Nowak, F. Jurie and B. Triggs. *Sampling Strategies for Bag-of-features Image Classification*. In Proceedings of the European Conference on Computer Vision (ECCV), 2006. 17

[Ojala *et al.* 2002] T. Ojala, M. Pietikainen and T. Maenpaa. *Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 24, no. 7, pages 971–987, 2002. 16

[Olshausen & Field 1997] B. A. Olshausen and D. J. Field. *Sparse Coding with An Overcomplete Basis Set: A Strategy Employed by V1?* Vision research (VR), vol. 37, no. 23, pages 3311–3325, 1997. 18

[Oquab *et al.* 2014] M. Oquab, L. Bottou, I. Laptev and J. Sivic. *Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 39

[Oquab *et al.* 2015] M. Oquab, L. Bottou, I. Laptev and J. 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 (CVPR), 2015. 39

[Otsu 1979] N. Otsu. *A Threshold Selection Method from Gray-Level Histograms*. IEEE Transactions on Systems, Man, and Cybernetics (TSMC), vol. 9, no. 1, pages 62–66, 1979. 51

[Ouyang *et al.* 2015] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C. C. Loy and X. Tang. *DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 58, 59

[Pandey & Lazebnik 2011] M. Pandey and S. Lazebnik. *Scene Recognition and Weakly Supervised Object Localization with Deformable Part-Based Models*. In Proceedings of the International Conference on Computer Vision (ICCV), 2011. 37, 39, 46, 52, 57, 61, 62, 63, 64, 65, 66, 67, 72, 78, 79, 110

[Park *et al.* 2000] D. K. Park, Y. S. Jeon and C. S. Won. *Efficient Use of Local Edge Histogram Descriptor*. In Proceedings of the ACM Workshops on Multimedia (MM-W), 2000. 15

[Pass *et al.* 1996] G. Pass, R. Zabih and J. Miller. *Comparing Images Using Color Coherence Vectors*. In Proceedings of the ACM International Conference on Multimedia (MM), 1996. 15

[Pedersoli *et al.* 2011] M. Pedersoli, A. Vedaldi and J. Gonzàlez. *A Coarse-to-fine Approach for Fast Deformable Object Detection*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011. 28

[Pennington *et al.* 2014] J. Pennington, R. Socher and C. D. Manning. *GloVe: Global Vectors for Word Representation*. In Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014. 91

[Perronnin *et al.* 2010] F. Perronnin, J. Sánchez and T. Mensink. *Improving the Fisher Kernel for Large-scale Image Classification*. In Proceedings of the European Conference on Computer Vision (ECCV), 2010. 72

[Pujol & Chen 2007] A. Pujol and L. Chen. *Line Segment Based Edge Feature Using Hough Transform*. In Proceedings of the International Conference on Visualization, Imaging and Image Processing (VIIP), 2007. 15

[Redmon *et al.* 2016] J. Redmon, S. Divvala, R. Girshick and A. Farhadi. *You Only Look Once: Unified, Real-Time Object Detection*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 82, 112

[Ren & Ramanan 2013] X. Ren and D. Ramanan. *Histograms of Sparse Codes for Object Detection*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013. 26, 46

[Ren *et al.* 2015a] S. Ren, K. He, R. Girshick and J. Sun. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. In Proceedings of Advances in Neural Information Processing Systems (NIPS), 2015. 6, 9, 22, 24, 29, 30, 34, 35, 74, 82, 112

[Ren *et al.* 2015b] S. Ren, K. He, R. Girshick, X. Zhang and J. Sun. *Object Detection Networks on Convolutional Feature Maps*. arXiv preprint arXiv:1504.06066, 2015. 75

[Rochan & Wang 2015] M. Rochan and Y. Wang. *Weakly Supervised Localization of Novel Objects Using Appearance Transfer*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 40, 91, 93

[Rohrbach *et al.* 2010] M. Rohrbach, Michael S., György S., Iryna G. and B. Schiele. *What Helps Where – And Why? Semantic Relatedness for Knowledge Transfer*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010. 91

[Ronfard *et al.* 2002] R. Ronfard, C. Schmid and W. Triggs. *Learning to Parse Pictures of People*. In Proceedings of the European Conference on Computer Vision (ECCV), 2002. 17

[Rosenberg *et al.* 2005] C. Rosenberg, M. Hebert and H. Schneiderman. *Semi-Supervised Self-Training of Object Detection Models*. In IEEE Winter Conference on Applications of Computer Vision (WACV), 2005. 86

[Rothe & Schütze 2015] S. Rothe and H. Schütze. *AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes*. In the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL), 2015. 92, 97

[Russakovsky *et al.* 2015] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei. *ImageNet Large Scale Visual Recognition Challenge*. International Journal of Computer Vision (IJCV), vol. 0, no. 0, pages 1–42, April 2015. 4, 7, 8, 22, 23, 28, 31, 45, 55, 69, 72, 83, 85, 87

[Russakovsky 2015] O. Russakovsky. *Scaling Up Object Detection*. PhD thesis, Stanford University, 2015. 3

[Sánchez *et al.* 2013] J. Sánchez, F. Perronnin, T. Mensink and J. J. Verbeek. *Image Classification with the Fisher Vector: Theory and Practice*. International Journal of Computer Vision (IJCV), vol. 105, no. 3, pages 222–245, 2013. 18

[Schapire 2001] R. E. Schapire. *The Boosting Approach to Machine Learning An Overview*. In MSRI Workshop on Nonlinear Estimation and Classification, 2001. 20

[Schmid *et al.* 2000] C. Schmid, R. Mohr and C. Bauckhage. *Evaluation of Interest Point Detectors*. International Journal of Computer Vision (IJCV), vol. 37, no. 2, pages 151–172, 2000. 17

[Sermanet *et al.* 2014] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun. *OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks*. In International Conference on Learning Representations (ICLR), 2014. 22, 30, 58, 82

[Shao *et al.* 2015] L. Shao, F. Zhu and X. Li. *Transfer Learning for Visual Categorization: A Survey*. IEEE Transactions on Neural Networks and Learning Systems (TNNLS), vol. 26, no. 5, pages 1019–1034, 2015. 37, 39

[Shi *et al.* 2012] Z. Shi, P. Siva and T. Xiang. *Transfer Learning by Ranking for Weakly Supervised Object Annotation*. In Proceedings of the British Machine Vision Conference (BMVC), 2012. 40, 50, 69, 71

[Shi *et al.* 2013] Z. Shi, T. M. Hospedales and T. Xiang. *Bayesian Joint Topic Modelling for Weakly Supervised Object Localisation*. In Proceedings of the International Conference on Computer Vision (ICCV), 2013. 63, 64, 66, 69, 71

[Shu *et al.* 2015] X. Shu, G. Qi, J. Tang and J. Wang. *Weekly-Shared Deep Transfer Networks for Heterogeneous-Domain Knowledge Propagation*. In Proceedings of the ACM International Conference on Multimedia (MM), 2015. 40

[Simo-Serra *et al.* 2015] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua and F. Moreno-Noguer. *Discriminative Learning of Deep Convolutional Feature Point Descriptors*. In Proceedings of the International Conference on Computer Vision (ICCV), 2015. 19

[Simonyan & Zisserman 2015] K. Simonyan and A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. In International Conference on Learning Representations (ICLR), 2015. 23, 74, 75, 97, 104

[Singh *et al.* 2016] K. K. Singh, F. Xiao and Y. J. Lee. *Track and Transfer: Watching Videos to Simulate Strong Human Supervision for Weakly-Supervised Object Detection*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 40

[Siva & Xiang 2011] P. Siva and T. Xiang. *Weakly Supervised Object Detector Learning with Model Drift Detection*. In Proceedings of the International Conference on Computer Vision (ICCV), 2011. 36, 38, 69, 71, 72, 73

[Siva *et al.* 2012] P. Siva, C. Russell and T. Xiang. *In Defence of Negative Mining for Annotating Weakly Labelled Data*. In Proceedings of the European Conference on Computer Vision (ECCV), 2012. 38, 62, 63, 65, 66, 69, 71

[Sivic & Zisserman 2003] J. Sivic and A. Zisserman. *Video Google: A Text Retrieval Approach to Object Matching in Videos*. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2003. 18

[Song *et al.* 2011] Z. Song, Q. Chen, Z. Huang, Y. Hua and S. Yan. *Contextualizing Object Detection and Classification*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1585–1592, 2011. 58

[Song *et al.* 2014a] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui and T. Darrell. *On Learning to Localize Objects with Minimal Supervision*. In Proceedings of the International Conference of Machine Learning (ICML), 2014. 38, 72, 73

[Song *et al.* 2014b] H. O. Song, Y. J. Lee, S. Jegelka and T. Darrell. *Weakly-supervised Discovery of Visual Pattern Configurations*. In Proceedings of Advances in Neural Information Processing Systems (NIPS), 2014. 73

[Srebro *et al.* 2005] N. Srebro, J. D. M. Rennie and T. S. Jaakola. *Maximum-Margin Matrix Factorization*. In Proceedings of Advances in Neural Information Processing Systems (NIPS), 2005. 19

[Stricker & Orengo 1995] M. A. Stricker and M. Orengo. *Similarity of Color Images*. In Storage and Retrieval for Image and Video Databases III (SPIE), pages 381–392, 1995. 15

[Swain & Ballard 1991] M. J. Swain and D. H. Ballard. *Color Indexing*. International Journal of Computer Vision (IJCV), vol. 7, no. 1, pages 11–32, 1991. 15

[Szegedy *et al.* 2013] C. Szegedy, A. Toshev and D. Erhan. *Deep Neural Networks for Object Detection*. In Proceedings of Advances in Neural Information Processing Systems (NIPS). 2013. 5, 24, 30, 45, 58, 82

[Szegedy *et al.* 2015] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich. *Going Deeper with Convolutions*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 23

[Tang *et al.* 2014a] K. Tang, A. Joulin, L. J. Li and L. Fei-Fei. *Co-localization in Real-World Images*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 40, 41, 62, 63, 65, 66

[Tang *et al.* 2014b] Y. Tang, X. Wang, E. Dellandrea, S. Masnou and L. Chen. *Fusing Generic Objectness and Deformable Part-based Models for Weakly Supervised Object Detection*. In Proceedings of the IEEE International Conference on Image Processing (ICIP), 2014. 10, 50

[Tang *et al.* 2016a] Y. Tang, J. Wang, B. Gao, E. Dellandrea, R. Gaizauskas and L. 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 (CVPR), 2016. 10

[Tang *et al.* 2016b] Y. Tang, X. Wang, E. Dellandrea and L. Chen. *Weakly Supervised Learning of Deformable Part-Based Models for Object Detection via Region Proposals*. IEEE Transactions on Multimedia (TMM), vol. PP, no. 99, pages 1–1, 2016. 10

[Trulls *et al.* 2014] E. Trulls, S. Tsogkas, I. Kokkinos, A. Sanfeliu and F. Moreno-Noguer. *Segmentation-aware Deformable Part Models*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 27

[Turing 1950] A. M. Turing. *Computing Machinery and Intelligence*. Mind, vol. 59, no. 236, pages 433–460, 1950. 2

[Tuytelaars 2010] T. Tuytelaars. *Dense Interest Points*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010. 17

[Uijlings *et al.* 2013] J. Uijlings, K. van de Sande, T. Gevers and A. Smeulders. *Selective Search for Object Recognition*. International Journal of Computer Vision (IJCV), vol. 104, no. 2, pages 154–171, 2013. 6, 8, 29, 30, 31, 37, 38, 47, 49, 50, 87, 96, 112

[Vidal-Naquet & Ullman 2003] M. Vidal-Naquet and S. Ullman. *Object Recognition with Informative Features and Linear Classification*. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2003. 17

[Viola & Jones 2001] P. Viola and M. Jones. *Rapid Object Detection Using A Boosted Cascade of Simple Features*. In Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2001. 4, 17

[Wang *et al.* 2015] C. Wang, K. Huang, W. Ren, J. Zhang and S. Maybank. *Large-Scale Weakly Supervised Object Localization via Latent Category Learning*. IEEE Transactions on Image Processing (TIP), vol. 24, no. 4, pages 1371–1385, April 2015. 38, 58, 69, 71, 73, 74, 112

[Weber *et al.* 2000] M. Weber, M. Welling and P. Perona. *Unsupervised Learning of Models for Recognition*. In Proceedings of the European Conference on Computer Vision (ECCV), 2000. 21

[Winder *et al.* 2009] S. Winder, G. Hua and M. Brown. *Picking the Best DAISY*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. 17

[Xu *et al.* 2014] J. Xu, S. Ramos, D. Vázquez and A. M. López. *Domain Adaptation of Deformable Part-Based Models*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 36, no. 12, pages 2367–2380, 2014. 26

[Yan *et al.* 2014] J. Yan, Z. Lei, L. Wen and S. Z. Li. *The Fastest Deformable Part Model for Object Detection*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 28

[Yang & Ramanan 2013] Y. Yang and D. Ramanan. *Articulated Human Detection with Flexible Mixtures of Parts*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 35, no. 12, pages 2878–2890, 2013. 26

[Yang *et al.* 2013] Y. Yang, G. Shu and M. Shah. *Semi-supervised Learning of Feature Hierarchies for Object Detection in a Video*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013. 86

[Zeiler & Fergus 2014] M. D. Zeiler and R. Fergus. *Visualizing and Understanding Convolutional Networks*. In Proceedings of the European Conference on Computer Vision (ECCV), 2014. 23, 90

[Zhang *et al.* 2010] W. Zhang, Q. M. J. Wu, G. Wang and H. Yin. *An Adaptive Computational Model for Salient Object Detection*. IEEE Transactions on Multimedia (TMM), vol. 12, no. 4, pages 300–316, 2010. 45, 51

[Zhao & Pietikainen 2007] G. Zhao and M. Pietikainen. *Dynamic Texture Recognition Using Local Binary Patterns with An Application to Facial Expressions*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 29, no. 6, pages 915–928, 2007. 16

[Zhao *et al.* 2012] G. Zhao, T. Ahonen, J. Matas and M. Pietikainen. *Rotation-Invariant Image and Video Description With Local Binary Pattern Features*. IEEE Transactions on Image Processing (TIP), vol. 21, no. 4, pages 1465–1477, 2012. 16

[Zhou *et al.* 2010] X. Zhou, K. Yu, T. Zhang and T. S. Huang. *Image Classification Using Super-vector Coding of Local Image Descriptors*. In Proceedings of the European Conference on Computer Vision (ECCV), 2010. 18, 71, 74

[Zhou *et al.* 2015] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba. *Object Detectors Emerge In Deep Scene CNNs*. In International Conference on Learning Representations (ICLR), 2015. 84

[Zhu & Ramanan 2012] X. Zhu and D. Ramanan. *Face Detection, Pose Estimation, and Landmark Localization in the Wild*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012. 26

[Zhu *et al.* 2011] Y. Zhu, Y. Chen, Z. Lu, S.J. Pan, G.R. Xue, Y. Yu and Q. Yang. *Heterogeneous Transfer Learning for Image Classification.* In AAAI Conference on Artificial Intelligence (AAAI), 2011. 40

[Zhu *et al.* 2013] C. Zhu, C. E. Bichot and L. Chen. *Image Region Description Using Orthogonal Combination of Local Binary Patterns Enhanced with Color Information*. Pattern Recognition, vol. 46, no. 7, pages 1949–1963, 2013. 16

[Zhu *et al.* 2014] Y. Zhu, J. Zhu and R. Zhang. *Contextual Object Detection With Spatial Context Prototypes*. IEEE Transactions on Multimedia (TMM), vol. 16, no. 6, pages 1585–1596, 2014. 45

[Zitnick & Dollar 2014] L. Zitnick and P. Dollar. *Edge Boxes: Locating Object Proposals from Edges*. In Proceedings of the European Conference on Computer Vision (ECCV), 2014. 8, 29, 37, 50, 112