Progressive Visual Object Detection with Positive Training Examples Only

Ekaterina Riabchenko[1], Ke Chen[2], and Joni-Kristian Kämäräinen[2]

[1] Department of Mathematics and Physics, Lappeenranta University of Technology
[2] Department of Signal Processing, Tampere University of Technology

Abstract. Density-aware generative algorithms learning from positive examples have verified high recall for visual object detection, but such generative methods suffer from excessive false positives which leads to low precision. Inspired by the recent success of detection-recognition pipeline with deep neural networks, this paper proposes a two-step framework by training a generative detector with positive samples first and then utilising a discriminative model to get rid of false positives in those detected bounding box candidates by the generative detector. Evidently, the discriminative model can be viewed as a post-processing step which improves the robustness by distinguishing true positives from false positives that confuse the generative detector. We exemplify the proposed approach on public ImageNet classes to demonstrate the significant improvement on precision while using only positive examples in training.

Keywords: Object detection, generative learning, discriminative learning, Gabor features, Histogram of Oriented Gradients (HOG)

1 Introduction

The problem of visual object class detection has been actively investigated in the vision community for more than ten years. With the development of datasets (increasing the dataset size and introducing the challenging images largely varying in pose, scale, lighting conditions, etc.), a number of algorithms have been proposed to cope with such a problem. The first generation datasets, e.g., UIUC car dataset [1] or Caltech-4 [11], contained only a few classes with hundreds of examples, and almost perfect results were obtained with generative part-based models [6, 11]. These generative models were part-based describing appearance of local parts and tolerating their spatial distortion, but the most popular approaches were based on visual Bag-of-Words (BoW) [25], which omitted spatial structure of object parts and described the classes via their local part histograms. With the help of strong discriminative learning methods, the BoW approach obtained the top accuracy [4] on the second generation datasets, e.g., Caltech-101 [9]. However, discriminative approaches with BoW features failed when they were applied to objects with severe view point changes and occlusion, so the concept of explicit local parts to describe objects were resurrected. The Deformable

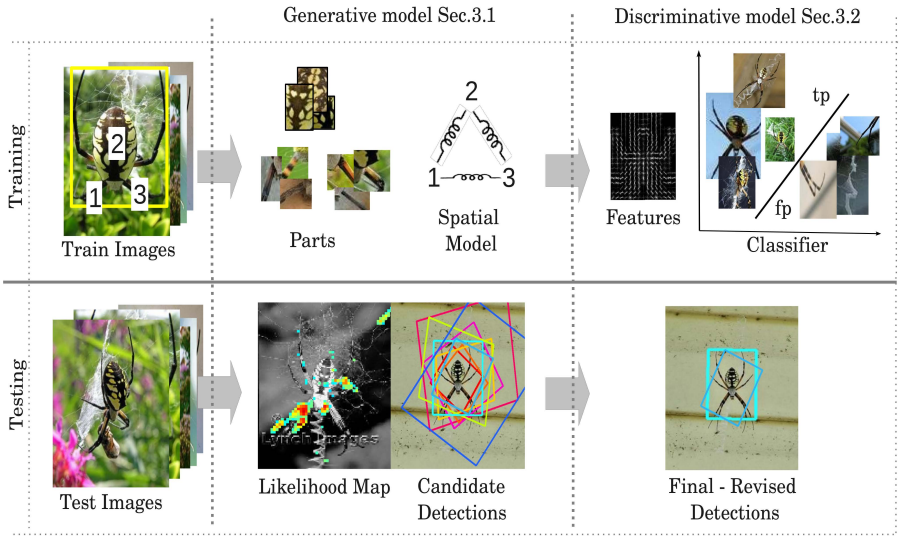


Fig. 1. Workflow of our proposed approach for visual object detection. To train a generative model, only positive instances of each object class with annotated bounding boxes and semantic object parts are employed. The true positive and false positive detections, obtained from generative model with training images, are used as positive and negative input examples of a discriminative model, which learns to discover their dissimilarities. During testing candidate object detections of the generative method are re-scored with the discriminative method, leading to the reduction of false positives and therefore the increase in precision. Here *tp* denotes true positives and *fp* - false positives.

Part Model (DPM) is the state-of-the-art discriminative part-based approach being constantly within the top performers in the third generation datasets, e.g., the annual Pascal VOC challenge [8]. Recently, the fourth generation datasets, such as ImageNet [24] and COCO [20], appear with thousands of classes and examples within them. The DPM model is clearly below the state-of-the-art [13, 29], but other discriminative model still dominate the field. In particular, the deep neural networks [16, 13] which have shown to implicitly learn a local part detector layers [14]. However, the complementary property of generative learning is that learning requires only positive examples. This leads to a large number of false positives which can be avoided by modelling the "no-class" (background) distribution to compute posterior probabilities for detection [9]. Detection fails if an ad hoc modelled background model or selected examples are poor.

To overcome the main limitation of the generative approach and still learn from positive examples only, this paper proposes a progressive generative-discriminative pipeline for object detection. The pipeline exploits the complementary properties of the two approaches: i) generative models first capturing the ap-

pearance distribution of a class and producing compact intra-class variance and ii) discriminative models learning the decision boundary between correct and false positives producing large inter-class variance. By separating these stages as compared to the existing monolithic systems, we can establish a generative-discriminative model for visual class detection (Fig. 1). Similar mechanism was introduced in Regions with Convolutional Neural Networks (RCNN) [13], which utilises a general objectness detector [2] to generate a large number of bounding box candidates and then applies the CNN classifier to obtain true positives in the images. Our system differs from their approach rather strongly in the sense that 1) our generative model produces much less candidates (i.e., two orders of magnitude reduction) and 2) requires only positive examples. Intuitively, our proposed framework can be viewed as a progressive coarse-to-fine pipeline: first localise the candidate locations with a generative object detector and then find true objects among those candidates with discriminative model. We exemplify our hypothesis using the recent bio-inspired fully probabilistic Generative Object Detector (GOD) [22, 23] and the state-of-the-art discriminative DPM model [10]. We demonstrate that superior precision-recall with challenging ImageNet classes can be achieved for the proposed two-stage pipeline.

2 Related Work

Visual Object Detection – The problem of object detection, which is to localise and classify objects appearing in still images, is a hot topic in computer vision. Due to challenges posed by variations in scale, pose, appearance, and lighting conditions, the problem attracts wide attention and a number of algorithms have been proposed. Existing object detection algorithms can be divided into two categories: model-free methods [3, 4, 13, 25, 29] and model-based methods [1, 6, 10, 11, 22, 23]. Specifically, the difference between model-free methods and model-based methods lie on the usage of the explicit object models with constraints between object parts. In the stream of model-free methods, the discrimination of feature representation plays a dominating role mitigating large variations of pose, scale and appearance, thus spatial pyramids of Bag-of-Words [17] and more recent deep features [13, 29] were adopted. On the other hand, with introducing the object models, both the appearance of local object parts and geometric correlation between object parts can be simultaneously learned in a unique framework. In earlier work of generative part-based constellation algorithms [11], location of parts was limited and only a sparse set of candidates, selected by a saliency detector, was considered. In [6], the proposed pictorial structure model can tolerate the changes of pose and geometric deformation of object, but label annotation for each object part was required. Alternatively, Riabchenko et al [22, 23] employed the bio-inspired Gabor features and Gaussian Mixture Models to capture both local appearance of parts and inter-part spatial correlation. Compared to the generative object detectors, the discriminative frameworks have also been proposed and achieved the superior detection performance. In discriminative multi-instance learning [10], the positive instances with

weak labels (i.e., only the bounding boxes for the whole object and no annotated labels for parts) and negative instances are used to localise both the object and also its parts with latent SVM.

Generative-Discriminative Approaches – Presented in [26] and described in the Sec. 1 advantages and disadvantages of discriminative and generative approaches in object recognition inspire researchers to develop a hybrid system to take the best of both paradigms. Specifically, hybrid generative-discriminative approaches have been widely proposed in different applications of computer vision such as scene classification [5], tracking [19] and image classification [18]. The hybrid approaches for visual object recognition can be divided into two categories: feature encoding based [18, 21] and learning based [12, 15, 28]. On one hand, in [21], generative part is used to encode a multi-level representation, which is then fed into a discriminative classifier for object recognition. On the other hand, most of the existing hybrid approaches [15, 28] incorporated discriminative classifier into a generative framework to improve the discrimination of the model. Framework in [12] shares similar structure as the proposed algorithm, but in our framework generative and discriminative stages are based on different and more generic features (Gabors and HOGs), while in [12] the same codebook representation is used by both stages of the hybrid system.

Contributions – The novelties and contributions of this paper are three-fold:

- The principle of the proposed generative-discriminative framework is a genetic paradigm for visual object detection, which generative and discriminative parts in our algorithm can be replaced readily by any detector.
- The observation that discriminative model can be utilised to improve the precision of generative detector is exploited in a coarse-to-fine manner.
- The experiments with the recent benchmarking ImageNet dataset verify the effectiveness of the proposed algorithm[3].

3 Methodology

As shown in Fig. 1, the pipeline of the proposed framework can be divided into generative part (see Sec. 3.1) and discriminative part (see Sec. 3.2). Sec. 3.3 presents the specifics of the generative-discriminative hybrid formulation.

3.1 Generative Object Detector (GOD)

The generative object detection, e.g., constellation model [11], is employed for object detection because of its capability of handling complex compositionality (large intra-class variations) [26]. In this section, we investigate the pipeline of Generative Object Detector (GOD) [22], which general structure is shown in Fig. 2. In generative part-based object detection algorithm, both bounding boxes containing the whole object and manually-labelled object parts are used to train the model. Therefore, the suggested part-based GOD employs both

[3] `https://bitbucket.org/EkaterinaRiabchenko/gabor_object_detector_code/`

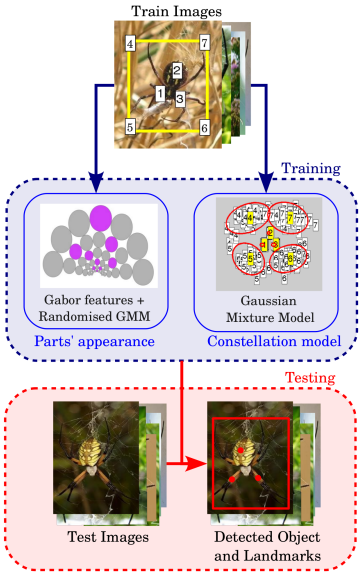


Fig. 2. The Generative Object Detector (GOD) framework for detecting visual classes.

local discriminative appearance of parts and also informative inter-part spatial arrangement.

To exclude the effect of geometric distortions on the object appearance model, all train images are aligned together prior to feature extraction. Images are aligned using homography transformation matching their parts' locations. In this aligned space object's structure becomes evident and is modelled along with the relative locations of the bounding box corners by the Gaussian mixture model. Each location is described with 2D Gaussian as illustrated in Fig. 2 (Constellation model). Object parts are modelled with the biologically inspired Gabor features, which have been successfully used in many vision applications. In order to reduce dimensionality of the features and provide a specifically optimised descriptor for each object part the concept of Randomised Gaussian Mixture Model is employed forming an appearance model [23].

During testing appearance model produces likelihood maps, which is then sampled for candidate locations of object parts with consecutive suppression procedure. The final step is the search for feasible object hypothesis (the required number of hypothesis can be predetermined), when candidate locations are pruned using constellation model and prior information about data statistics (Fig. 1 - Testing). Nevertheless, pruning still keeps a lot of false positive detections what results in relatively low precision in the presence of the high recall. This observation and the fact that GOD scores are likelihoods - not posterior

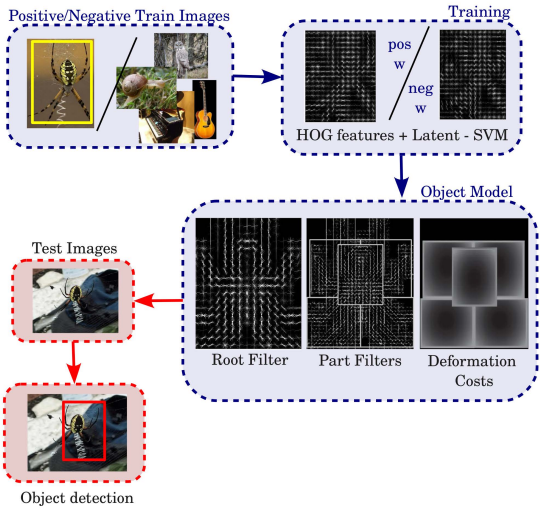


Fig. 3. The Discriminative Object Detector (DOD), e.g., deformable part-based model (DPM), for learning and detecting visual categories. Note that in our final system the negative training examples are produced by the GOD detector.

probabilities of object vs. non-object detection encourage us to add a discriminative part to solve the problem.

3.2 Discriminative Object Detector (DOD)

The discriminative models of visual class detection can be divided to scanning window and part-based models. A successful and popular example of the scanning window is the Viola-Jones detector [27]. The strength of the scanning window approach is conceptual simplicity, but the drawback is difficulty to capture distortions by view point change, occlusion and background clutter. Moreover, an effective scanning window method requires a large number of training examples.

Robustness to the distortions can be improved by dividing the window to sub-window bins and computing a histogram feature where the histogram dimensions represent the spatial structure as binned features [17]. This approach is adopted in the Histogram of Oriented Gradients (HOG) [7]. Local parts of objects are typically sufficiently rigid and therefore HOGs can capture them well. HOGs and HOG based deformable part-based model (DPM) [10] are used in our experiments as discriminative stage. The DPM has only a few tunable parameters, owing it to the fact that selection of the parts, learning their descriptors and learning the discriminative function for detection are all embedded in the latent support vector machine framework (see Fig. 3). Intuitively, DPM model is to alternatingly optimise the learning weights and the relative locations of deformable part filters in order to achieve high response in foreground and

low response in background. With the learned DPM model, the part learned filters are used to scan the whole feature pyramid map to find regions with high response, which can finally determine locations of the object.

3.3 Hybrid Generative-Discriminative Pipeline

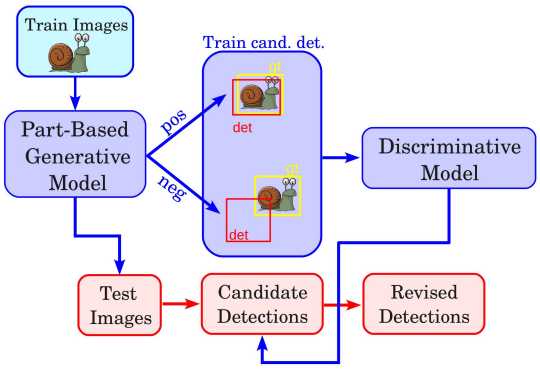


Fig. 4. Generative-discriminative hybrid approach.

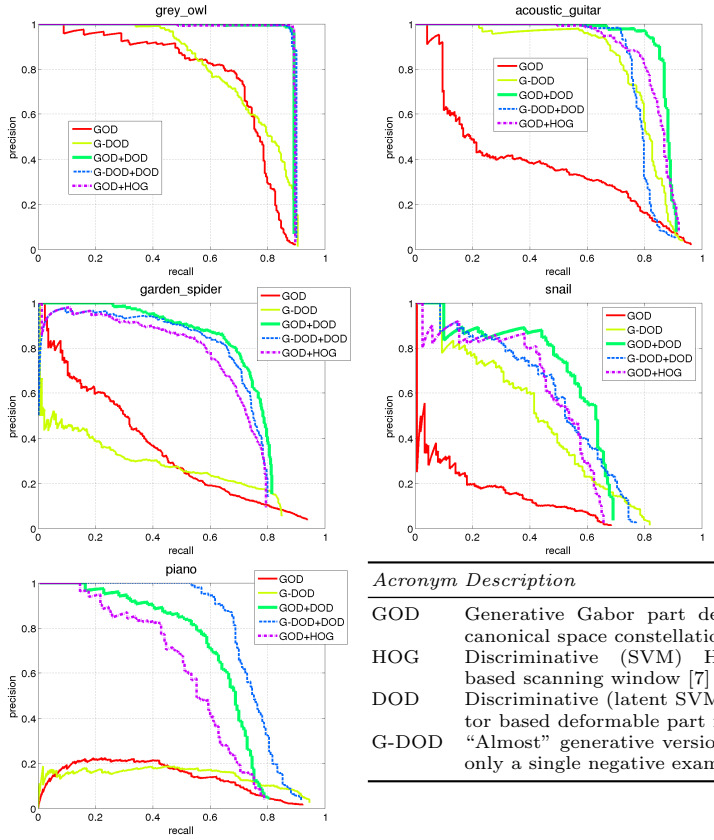
The limitations of using either generative or discriminative approaches solely encourages us to develop a hybrid detector to overcome the limitations. In related work (Sec. 2) state-of-the-art hybrid approaches were described, encouraging us to adopt a two-stage pipeline of a discriminative object detector to improve the detection performance of a generative one. In this pipeline the discriminative object detector can be viewed as a post-processing stage of the generative object detector. Fig. 4 illustrates the pipeline of the proposed hybrid method.

The generative method is trained by positive examples of the query image category with annotated object parts and bounding boxes. Discriminative method, on the other hand, uses the training stage outputs of the generative method as its inputs. Candidate detections of the generative method are transformed to the aligned space using the detected part locations. After alignment detections are scaled to the size 64×64 pixels and subsequently fed to the discriminative part for re-scoring. Generative output candidates having bounding box overlap ratio $A > 0.7$ with the ground truth are used as positive examples in discriminative training and the outputs with $A < 0.2$ as the negatives. This representation of positive and negative data allows the discriminative method to learn, exploit and emphasise the difference in appearance of the true positives and false positives that the generative part produces but is blind itself [13]. During testing the discriminative re-scoring procedure is applied to all detection candidates produced by generative method on the test data. These re-scored detections of the

discriminative stage are further processed by non-maximum suppression which removes spatially overlapping candidates. Non-maximum suppression removes candidates hyp_i with lower scores if their overlap ratio is greater than 0.5:

$$\frac{B_{hyp1} \cap B_{hyp2}}{B_{hyp1} \cup B_{hyp2}} \geq 0.5 \ . \tag{1}$$

4 Experiments



Acronym	*Description*
GOD	Generative Gabor part detector [23] and canonical space constellation model [22]
HOG	Discriminative (SVM) HOG descriptor based scanning window [7]
DOD	Discriminative (latent SVM) HOG descriptor based deformable part model [10]
G-DOD	"Almost" generative version of DOD using only a single negative example.

Fig. 5. Precision-recall curves for the Imagenet categories *grey owl*, *acoustic guitar*, *garden spider*, *snail* and *piano*. All methods use generative and use only positive examples in training.

4.1 Settings

Five challenging categories from the ImageNet database were used in our experiments: acoustic guitar, piano, snail, garden spider and grey owl. Categories are represented with objects appearing in different scales, orientations, lighting conditions, with limited 3D pose changes and moderate intra-class variation and were randomly divided into training and testing groups of approximately the same size.

4.2 Performance Metrics

A widely accepted measure for the object detection accuracy, overlap ratio, is used in our work. Overlap ratio A reflects precision of bounding box based object localisation. Object is considered to be located correctly if $A > 0.5$, where A is equal to intersection/union ratio of the ground truth and candidate (hypothesis) detection:

$$A = \frac{B_{gt} \cap B_{hyp}}{B_{gt} \cup B_{hyp}} \ . \tag{2}$$

In general, a generative method can produce a large number of hypotheses to guarantee that at least one passes the test in (2). That would result high recall, but poor precision which is the problem of generative methods. The discriminative part of the proposed pipeline aims to reduce the number of *false positives* (fp) by keeping the number of *true positives* (tp) high. Therefore it is meaningful to compute precision and recall values in the following way:

$$Precision = \frac{tp}{tp + fp}, \quad Recall = \frac{tp}{\text{Total number of positives}}.$$

4.3 Results

The results of various implementations of the proposed pipeline are shown in Fig. 5. The implementations are based on the publicly available code: Gabor object detector (GOD) by the authors [22, 23], scanning window based detector (HOG) by Dalal and Triggs [7], and the state-of-the-art discriminative part-based model (DOD) by Felzenszwalb et. al. [10]. In addition to the standard DOD we constructed a generative version of it (G-DOD) by allowing only a single negative example. From the results in Fig. 5 it is obvious that the plain generative methods (GOD, G-DOD) achieve high recall but poor precision - they detect the correct class, but are also triggered by many other things. The tested progressive generative-discriminative combinations (GOD+DOD, G-DOD+DOD, GOD+HOG) achieve almost the same recall as the generative, but with significantly better precision. The two strongest combinations are GOD+DOD and G-DOD+DOD indicating the superiority of the part-based approaches.

5 Conclusions

This paper proposes a hybrid generative-discriminative learning paradigm for visual class detection - a paradigm where the generative part generates detection candidates and the discriminative part learns to separate true and false positives. In the experiments, the paradigm significantly improved performance of generative detectors which is explained by the two-stage system where the discriminative stage mitigates the presence of excessive false positives. Our results indicate that despite the huge popularity of the discriminative learning in visual class detection, also the generative approach without ad hoc background modelling can be adopted, but needs to be paired with the discriminative true positive classifier. Our results show that this can be achieved without sacrificing the main properties of generative learning: training with positive examples only and presence of other classes does not affect the trained detector. We believe that the proposed framework can be facilitated in large scale visual problems where the best part-based discriminative methods (e.g., [10]) fail due to ambiguity between fine-grained class differences [29].

ACKNOWLEDGEMENTS

This work is funded by Academy of Finland under the grant no. 267581.

References

1. Agarwal, S., Roth, D.: Learning a sparse representation for object detection. In: European Conference on Computer Vision, ECCV 2002, pp. 113–127. Springer (2002)
2. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(11), 2189–2202 (2012)
3. Blaschko, M.B., Lampert, C.H.: Learning to localize objects with structured output regression. In: European Conference on Computer Vision, ECCV 2008. pp. 2–15. Springer (2008)
4. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: IEEE International Conference on Computer Vision, ICCV 2007. pp. 1–8 (2007)
5. Bosch, A., Zisserman, A., Muoz, X.: Scene classification using a hybrid generative/discriminative approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(4), 712–727 (April 2008)
6. Crandall, D., Felzenszwalb, P., Huttenlocher, D.: Spatial priors for part-based recognition using statistical models. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005. vol. 1, pp. 10–17 (2005)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005. vol. 1, pp. 886–893 (2005)

8. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html (2012)
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding 106(1), 59–70 (2007)
10. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1627–1645 (2010)
11. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2003. vol. 2, pp. 264–271 (2003)
12. Fritz, M., Leibe, B., Caputo, B., Schiele, B.: Integrating representative and discriminant models for object category detection. In: IEEE International Conference on Computer Vision, ICCV 2005. vol. 2, pp. 1363–1370 (2005)
13. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Computing Research Repository arXiv.org, CoRR. vol. abs/1311.2524 (2013)
14. Girshick, R.B., Iandola, F.N., Darrell, T., Malik, J.: Deformable part models are convolutional neural networks. CoRR abs/1409.5403 (2014), `http://arxiv.org/abs/1409.5403`
15. Kapoor, A., Winn, J.: Located hidden random fields: Learning discriminative parts for object detection. In: European Conference on Computer Vision, ECCV 2006, pp. 302–315. Springer (2006)
16. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: NIPS (2012)
17. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006. vol. 2, pp. 2169–2178 (2006)
18. Li, Y., Shapiro, L.G., Bilmes, J.A.: A generative/discriminative learning algorithm for image classification. In: IEEE International Conference on Computer Vision, ICCV 2005. vol. 2, pp. 1605–1612 (2005)
19. Lin, R.S., Ross, D., Lim, J., Yang, M.H.: Adaptive discriminative generative model and its applications. In: Advances in neural information processing systems, NIPS 2004. pp. 801–808. The MIT Press (2004)
20. Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision, ECCV 2014. Springer (2014)
21. Perina, A., Cristani, M., Castellani, U., Murino, V., Jojic, N.: A hybrid generative/discriminative classification framework based on free-energy terms. In: IEEE International Conference on Computer Vision, ICCV 2009. pp. 2058–2065 (2009)
22. Riabchenko, E., Kämäräinen, J.K., Chen, K.: Density-aware part-based object detection with positive examples. In: International Conference on Pattern Recognition, ICPR 2014. IEEE (2014)
23. Riabchenko, E., Kämäräinen, J.K., Chen, K.: Learning generative models of object parts from a few positive examples. In: International Conference on Pattern Recognition, ICPR 2014. IEEE (2014)

24. Russakovsky, O., Deng, J., Huang, Z., Berg, A.C., Fei-Fei, L.: Detecting avocados to zucchinis: what have we done, and where are we going? In: IEEE International Conference on Computer Vision, ICCV 2013. pp. 2064–2071 (2013)
25. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: IEEE International Conference on Computer Vision, ICCV 2003. pp. 1470–1477 (2003)
26. Ulusoy, I., Bishop, C.M.: Generative versus discriminative methods for object recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005. vol. 2, pp. 258–265 (2005)
27. Viola, P., Jones, M.: Robust real-time face detection. International journal of computer vision 57(2), 137–154 (2004)
28. Zhang, D.Q., Chang, S.F.: A generative-discriminative hybrid method for multi-view object detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006. vol. 2, pp. 2017–2024 (2006)
29. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: European Conference on Computer Vision, ECCV 2014. pp. 834–849. Springer (2014)