Unsupervised Object Detection for Automatic Video Annotation

Aneri Manoj Gandhi
School of Electrical Sciences
Indian Institute of Technology, Bhubaneswar

***Abstract*—We propose a pipeline that helps solve insufficient existing, fully labeled categories for object detection. The model suggested assists automatic annotation of unencountered objects in videos provided with a general label. In the framework, we first learn to extract all the possible objects present, i.e., to learn the objectness using a `Mask_RCNN` based binary classifier. The knowledge would generalize well to new categories and could assist detection models. The features are then extracted from the RoI proposed by the classifier and clustered. Based on clustering output and certain assumptions, the object described by the general label is identified.**

***Index Terms*—Object Detection, Unsupervised learning, Clustering.**

I. Introduction

Object detection, as one of the most fundamental and challenging problems in computer vision, has received significant attention in recent years. It has drastically improved in performance and scale with the development of deep convolution neural networks (CNNs) [1], [2], [3], [4]. Its development in the past two decades is regarded as the epitome of computer vision history (Fig. 1).

Most state-of-the-art models rely on supervised learning algorithms, requiring a vast quantity of fully labeled images with bounding box annotations. Thus, it is not feasible for hundreds and thousands of new object classes to obtain images with bounding box annotations in the real world. However, obtaining weakly labeled images and videos is relatively easy and can be obtained through a simple Google or YouTube search.

Through this project, we aim to design a pipeline that helps in automatic bounding box annotation of videos, which trivially extends to images. Leveraging the concepts of image pyramids, we contemplate that at higher levels, the objectness for a variety of classes can be realized using a simple binary classifier that differentiates between object and non-object. For example, the high-level features of a box and a Rubik's cube are very similar (edges, planes, corners, etc.). Hence, if a model can detect a box as an object, it certainly classifies Rubik's cube as an object.

Building upon the above idea, once an object is detected, its features can be extracted. A video (or an image) contains multiple objects other than its general label/ topic. These extra set of objects act as "distractors" for original object to be detected. Hence, it is necessary to match the object and its label correctly. We have assumed that the object we wish to label with the given label is present in most of the videos' frames. Using the assumption and the clusters made, we can then try to obtain what is required.

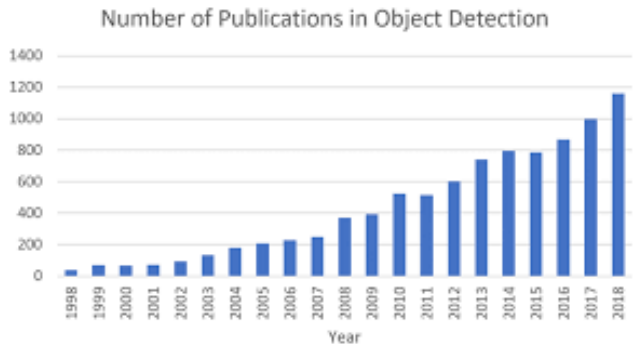

Fig. 1. The increasing number of publications in object detection from 1998 to 2018. (Data from Google scholar advanced search: allintitle: "object detection" AND "detecting objects".)

II. Related works

The project combines many fields of study on which extensive research is being done around the globe.

A. Object detection

As Fig. 1 suggests there has been a lot of work in the field of object detection. Most of the early object detection algorithms were built based on handcrafted features.

Introduction of RCNN by R. Girshick et al. acted as catalyst in development in object detection [5].Fast-RCNN and Faster-RCNN further improves upon RCNN by simultaneously training a detector and a bounding box regressor under the same network configurations. T.-Y. Lin et al. proposed Feature Pyramid Networks (FPN) on basis of Faster RCNN [6]. A topdown architecture with lateral connections was developed in FPN for building high-level semantics at all scales. It showed great advances for detecting objects with a wide variety of scales.Mask RCNN [7] adds pixel level masks to the image on top of Faster-RCNN. Most on these object detection algorithm depend on availability of fully annotated data of object class to be detected. We use object detection models to abstract objectness and then build model to automatically detect and annotate objects.

B. Unsupervised learning

Unsupervised learning poses one of the most difficult challenges in computer vision today. This field is being researched by many as it has close correlation with the way humans perceive and recognize images.However, most of the current models are based on supervised learning algorithm. In one of the paper that does multi-object recognition and classification [8], variance and gradient variance are compressed based on

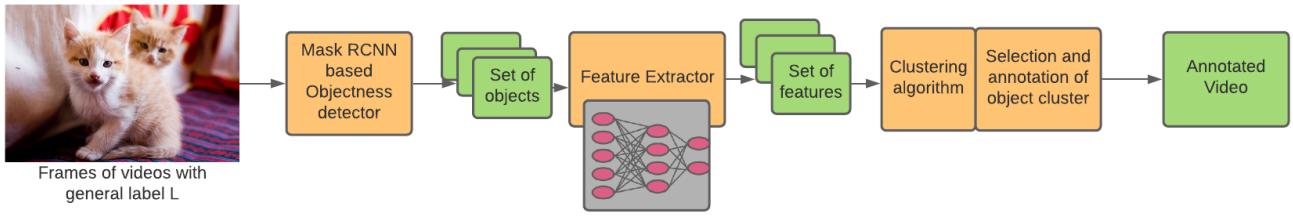

Fig. 2. Overview of the pipeline being followed.

PCA (Principal Component Analysis) to get initial classifications of clusters via K-means algorithm in the image frame. Nucleus of each cluster is identified and using cell expansion object is recognized and labeled. Unlike them, we have considered a deep neural network architecture. One of the study [9] has adopted an approach where the unsupervised learning phase is during training time, then at test time standard feed-forward processing is applied. We train the objectness module with supervised algorithm and the further use unsupervised algorithm.

C. Objectness transfer

Learning objectness to enhance detection performance has been studied by many previous works [10], [11], [12].Many of them explicitly train it to distinguish objects with a well-defined boundary in space and measure objectness with low level cues such as saliency [10], while others uses contour information and hierarchical superpixels.The ability to extract objectness knowledge is constrained as these papers do not consider the similarities in high level features as described in the introduction. These high level features can prove extremely beneficial for detecting new objects. Based on this consideration, learning objectness from CNN-based objectness knowledge has already been explored by DeepBox [11]. Work on robust objectness transfer has also been done where objectness is learned using mixed supervision [1].Objectness is learned using strong categories (bounding box annotated images) and weak categories (image level labelling). Strong categories are used to learn objectness and combined both provide domain invariance.These paper aim for improved object detection using objectness unlike us, where we aim to annotate objects in videos using objectness.

III. Proposed Work

The problem addressed here is to give bounding box annotation for object class given a weak label of videos by using an unsupervised learning algorithm. Elaborating, given set of videos (S) with a label (L), identify and object o from the set of objects (O) present in S, which is referred by L. Thus, the problem can be divided into three sub-parts:

- Object extraction or objectness detection
- Feature extraction
- Clustering and Object identification

IV. Methodology

The pipeline is summarized in Fig. 2. We will be looking into each module in detail in the following sections. Section A describes how we would obtain the set of objects O from a given video or set of images. Section B summaries the proposed methodology to extract features from O and finally the module three will be explained in section C.

A. Objectness Detection

For this module, object extraction, we aim to model the objectness knowledge using a CNN-based method.

The data set used is the standard PASCALVOC(2012) dataset. As a part of preprocessing, the ground truth annotations were changed to ”object” from specific class annotations like ”person,” ”cycle,” etc.

Leveraging the bounding box annotations of the available dataset and considering the regions that largely overlap with the ground truth boxes as the ”objects” and the regions with smaller overlaps as ”non-objects”. We aim to let CNN models figure out automatically the helpful cues for learning ”objectness” [1].

The initial attempt was to use a ResNet-101 classifier [13] and a bounding box regressor (a small variant of R-CNN) [5]. Due to some dependency issues and resource constraints, an alternate model was used. The classifier used was Resnet-101 only. However, considering the requirements, Feature Pyramid Network (FPN) [6]was expected to show promising results as its idea of detecting objects at different scales inlined with the concept we wish to implement to extract ”objectness.” Hence we use Mask-RCNN based on ResNet-101 and FPN. Firstly, we trained the CNN for classification. We trained a new model starting from weights of Resnet pre-trained on the COCO dataset.The maximum image size considered is 1024x1024x3 pixels. If an input image is smaller than that, it is wrapped to obtain the required size. Fig. 3 shows a general M-RCNN network.

The region proposal network is then fine-tuned.Positive samples have IoU (intersection-over-union) > 0.7, while negative samples have IoU < 0.3.The non-maxima-suppression threshold is considered as 0.3.

The multi-task loss function of Mask R-CNN combines the loss of classification, localization and segmentation mask:

$$L = L_{cls} + L_{box} + L_{mask} \quad (1)$$

The loss function sums up the cost of classification, bounding box prediction and segmentation mask [14]. Individually, the losses are calculated as:

$$L_cls = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) \quad (2)$$

$$L_{cls}(p_i, p_i^*) = -p_i^* log(p_i) + (1 - p_i^*) log(1 - p_i) \quad (3)$$

$$L_{box} = \frac{\lambda}{N_{box}} \sum_i p_i^* L_1^{smooth}(t_i - t_i^*) \quad (4)$$

$$L_mask = \frac{-1}{m^2} \sum_{1 \le i,j \le m} [y_{ij} log(\hat{y}_{ij}^k) + (1 - y_{ij}) log(1 - \hat{y}_{ij}^k)] \quad (5)$$

where $p_i$is predicted probability of anchor i beign object, p_i^* is ground truth, t_i is four parametrized coordinates and t_i^* are ground truth coordinates. All Ns are normalization terms and y_{ij} is predicted value of mxm mask.

The model has been trained and tested on the test set provided by PASCALVOC. We will further test the model on and on objects absent in the standard datasets in YouTube videos.

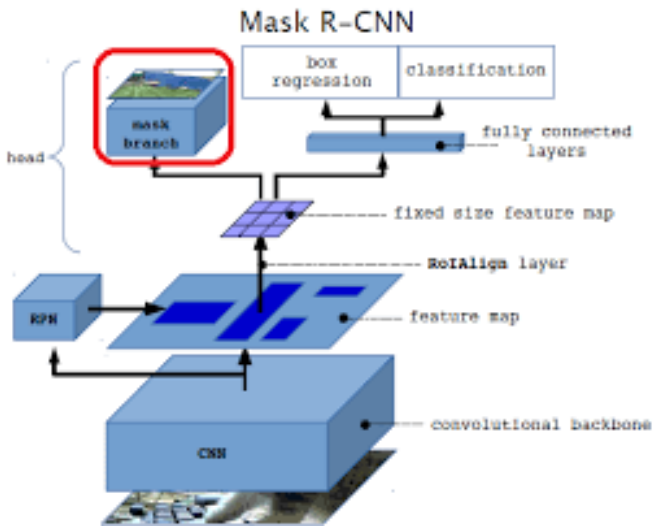

Fig. 3. General structure of M-RCNN. Here, FPN classif fc layers used is 1024, Anchors considered per image is 256 and mask of size 28x28 is used.

B. Feature Extraction

From the set of objects extracted, we now want to extract distinct features. The main challenge of this module is to differentiate objects with similar features. Since the module one does not distinguish between objects, extracting features as vectors that discriminate between similar objects, such as a box and Rubik's cube, without a supervised algorithm is complicated. Output required from this module are feature vectors for all objects extracted. Both image descriptors and neural networks can obtain the vectors. The image descriptor algorithms are based on gradient descent across the image. Most famous of them are SURF, ORB, SIFT, BRIEF. For neural networks, Auto-encoders are being used for feature extraction. Auto-encoder is an unsupervised learning technique in which we leverage neural networks for the task of representation learning. However, they are lossy, and due to its dimensionality reduction process, it tends to capture features common to many classes. Convolution Neural Networks are also commonly used for extracting features. There are few networks like Dilated Residual networks (DRN) that extract minute feature and provide high-resolution feature maps [15]. Due to DRN's proposed accuracy and performance in downstream applications, it seems to be the best option. However, we would explore all the options, and a precise decision is not yet made.

C. Clustering and identification

The feature vectors obtained from the previous module act as a big set of unlabeled data. Many instances of a single object are expected to be present in the set. Hence, the first step is to identify those same objects. Identification can be easily implemented using clustering them. There are multiple clustering algorithms with no ordering for the best algorithm. Some clustering algorithms require us to specify or guess the number of clusters to discover in the data. In contrast, others require the specification of some minimum distance between observations in which examples may be considered "close" or "connected." Depending upon the nature of feature vectors obtained, the algorithm to be used can be selected.

We have assumed here that number of instances of object described by the video title would be considerably more compared to other objects present. Hence, we reduce our search set from all object set obtained to top five object cluster which are most dense. To select from these 5, selection criteria such as an gap between density will be implemented. There will be certain objects like person class that would be present in most cases in the top 5 but may not be required. These classes can then be segregated further. Finally, we can backtrack and locate the instances of object and annotate them with bounding box and labels.

V. Results

Fig. 4 displays few output images obtained from testing the objectness model on PASCALVOC test data. The actual accuracy of the objectness detector can only be calculated for object classes it is not trained on. Obtaining ground truth annotated test set. Hence, for mathematical evaluation of the model, a test set will be created. From general observation we can conclude that its performance is acceptable.In Fig. 4.a, we can see an extra object segmented with yellow cover at the same time we observe that not all computers are extracted as objects in Fig. 4.e . However, considering a video, where two adjacent frames has almost same components, average error reduces.

VI. Future Work

Also the Module one will be evaluated and tested against new objects following which Module two and three will be implemented. Various techniques for feature extraction will

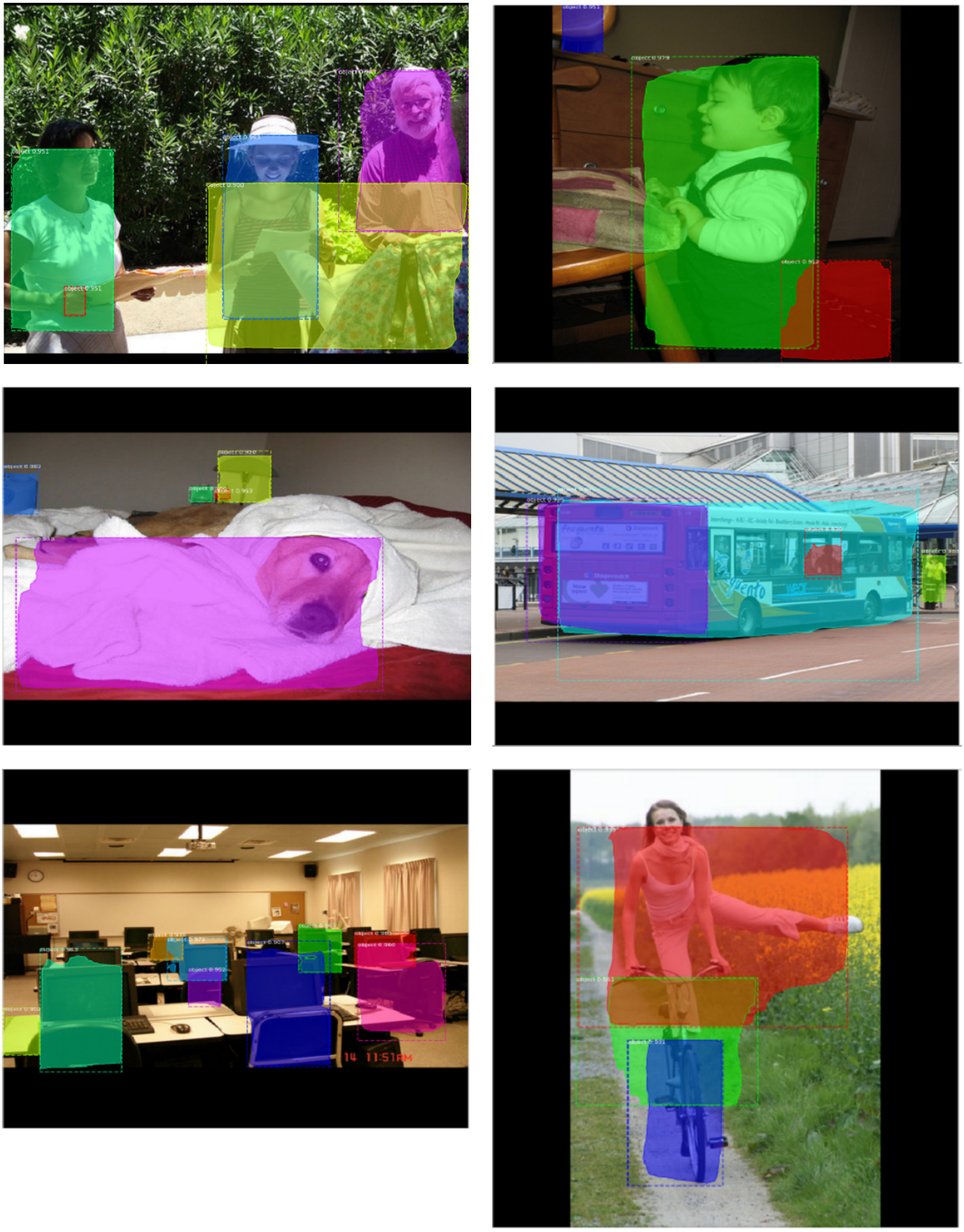

Fig. 4. Output of Objectness Detector on test set.

be studied and most suitable will be implemented. Depending upon the results of first two modules, clustering algorithm will be chosen and implemented. We believe we can further improve the accuracy of the model in future once entire pipeline is ready.

VII. Conclusion

Thus, through this project we propose an architecture that would help to annotate objects in videos without any teacher supervision. We can generalize this model and use it to annotate set of images with common image level label. The objectness transfer ensures that the pipeline stands for any out of the ordinary object class. It can help generate train and test dataset for the same.

By working on this project, we were able to gain insights about deep convolution neural networks like Resnets, DRN and MRCNN. Also implementing them enabled me to understand their learning process.

Acknowledgment

I would like to thank my supervisor Dr. Debi Prasad Dogra for guiding me through out the period and encouraging me to explore enabling me gain knowledge in the field of computer vision.

References

[1] Y. Li, J. Zhang, K. Huang, and J. Zhang, "Mixed supervised object detection with robust objectness transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 639–653, 2019.

[2] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[3] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[4] P. Burlina, "Mrcnn: A stateful fast r-cnn," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 3518–3523.

[5] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013. [Online]. Available: http://arxiv.org/abs/1311.2524

[6] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *CoRR*, vol. abs/1612.03144, 2016. [Online]. Available: http://arxiv.org/abs/1612.03144

[7] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870

[8] R. Luo, P.-Y. Chuang, and X.-Y. Yang, "Multi-objects recognition using unsupervised learning and classification," 05 2013, pp. 1–6.

[9] I. Croitoru, S. Bogolin, and M. Leordeanu, "Unsupervised learning of foreground object detection," *CoRR*, vol. abs/1808.04593, 2018. [Online]. Available: http://arxiv.org/abs/1808.04593

[10] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012.

[11] W. Kuo, B. Hariharan, and J. Malik, "Deepbox: Learning objectness with convolutional networks," *CoRR*, vol. abs/1505.02146, 2015. [Online]. Available: http://arxiv.org/abs/1505.02146

[12] L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*, September 2014. [Online]. Available: https://www.microsoft.com/en-us/research/publication/edge-boxes-locating-object-proposals-from-edges/

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[14] L. Weng, "Object detection for dummies part 3: R-cnn family," *lilianweng.github.io/lil-log*, 2017. [Online]. Available: http://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html

[15] F. Yu, V. Koltun, and T. A. Funkhouser, "Dilated residual networks," *CoRR*, vol. abs/1705.09914, 2017. [Online]. Available: http://arxiv.org/abs/1705.09914