Improving Object Detection Performance Using Scene Contextual Constraints

Journal:	*IEEE Transactions on Cognitive and Developmental Systems*
Manuscript ID	TCDS-2019-0292.R2
Manuscript Type:	SI: Development and Learning and on Epigenetic Robotics
Date Submitted by the Author:	n/a
Complete List of Authors:	Alamri, Faisal; University of Exeter, Computer Science Pugeault, Nicolas; University of Exeter College of Engineering Mathematics and Physical Sciences, Computer Science
Keywords:	Neural network, object detection, contextual information, out-of-context, relabelling, Faster RCNN, visual system and development, cognitive system and development

Improving Object Detection Performance Using Scene Contextual Constraints

Faisal Alamri, Nicolas Pugeault, *Member, IEEE,*

(Invited Paper)

***Abstract*—Contextual information, such as the co-occurrence of objects and the spatial and relative size among objects, provides rich and complex information about digital scenes. It also plays an important role in improving object detection and determining out-of-context objects. In this work, we present contextual models that leverage contextual information (16 contextual relationships are applied in this paper) to enhance the performance of two of the state-of-the-art object detectors (*i.e.*, Faster RCNN and YOLO), which are applied as a post-processing process for most of the existing detectors, especially for refining the confidences and associated categorical labels, without refining bounding boxes. We experimentally demonstrate that our models lead to enhancement in detection performance using the most common dataset used in this field (MSCOCO), where in some experiments PASCAL2012 is also used. We also show that iterating the process of applying our contextual models also enhances the detection performance further.**

***Index Terms*—Neural network, object detection, contextual information, re-scoring, relabelling, out-of-context, semantic, spatial, scale.**

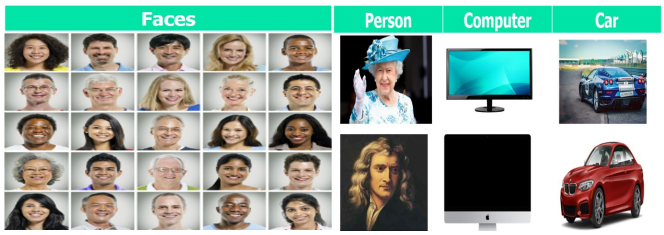
I. Introduction

HOW do we interpret visual scenes? A mere glance at an image is usually sufficient for us to recognise what objects compose the scene, and to understand its contents. This task remains challenging for computer vision systems, in spite of rapid improvements in the field, lead in particular by the development of deep learning approaches [47].

Object detection aims to determine whether an object for some predefined classes is present in some given images, where it outputs both the *what* and the *where* objects are in images. It is one of the most fundamental problems in the field of computer vision. Object detection has been used in a variety of applications such as sectary, robot vision, autonomous driving and scene captioning [30]. Object detection can be grouped into two categories, upon the aim of each type. First, *detection for single object*, which aims to detect an instance of an specific object class. However, the second type, known as *multiple object detection*, which seems to be more complex, aims to detect multiple objects of the pre-defined classes [30]. Figure 1 shows the differences between both types. Figure 1a shows different instances of the same object category, whereas in Figure 1b, it is shown different categories with two instances for each category, which is meant to present the latter detection type, as it is the main focus of the paper.

Faisal Alamri is with the Department of Computer Science, The University of Exeter, UK. E-mail: fa269@exeter.ac.uk

Dr. Nicolas Pugeault is supported by the Alan Turing Institute a Senior Lecturer in Computer Vision Machine Learning at the Univerity of Exeter E-mail: n.pugeault@exeter.ac.uk



(a) Single Object Detection (b) Object Category Detection

Fig. 1: Types of Object Detection

(a) Object isolated (b) Object in context

Fig. 2: Importance of Contextual Information

Contextual information plays an important role in visual recognition for both human and computer vision systems. Figure 2a shows an object isolated from its context, which seems hard to be identified not only by systems but even by some humans, whereas when presented in context (Figure 2b), it can be classified with less effort (*i.e.*, it is a *cup*). This example illustrates the fact that contextual information carries rich information about visual scenes. In terms of object recognition, it could be defined as cues captured from a scene that presents knowledge about objects locations, size and object-to-object relationships. Due to the importance of contextual information, it has been widely studied [6, 12, 18, 32, 37, 48].

In this paper, which is en extended version of our conference paper [1], we are proposing two models that leverage contextual information for re-scoring confidences in object detections and relabeling them when appropriate. Those models are applied as a post-processing process for most of the state-of-the-art detectors, for refining detected objects confidences, but without refining bounding boxes. The models obtain higher mAP, F1 and AUC scores compared to two of the state-of-the-art detectors (*i.e.*, Faster RCNN and YOLO). Some key features of the proposed models are:

1) They improve the detection performance of the state-of-the-art object detectors due to the inclusion of semantic, spatial and scale contexts, which are observed to be effective in this regards.

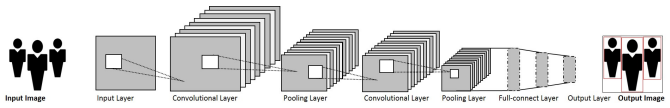


Fig. 3: Typical CNN Structure

2) They can be integrated as post-processing to most CNNs-based object recognition frameworks, whether detectors are one-stage or two-stages. This is in contrast to [41], which is specific to their end-to-end pipeline. See Section II for more details about CNNs-based detectors.
3) Rather than only evaluating whether the detected regions are correctly detected, eg, as in [7, 12], the proposed models also re-score and relabel detections.
4) They are defined from 16 contextual relationships, as presented in Section III, unlike other models that use a smaller number of relationships such as [3, 8, 12].
5) Compared to [13, 25, 37], which mainly re-score objects probabilities, the proposed relabeling model steps further to relabel objects and re-score them based on their relationships with other detected objects.

This paper is organised as follows: First, we review some of the state-of-the-art CNN-based detectors (Section II). The 16 contextual relationships proposed in this paper, including the datasets and methods applied for the proposed models are presented in Sections III and IV, respectively. Finally, the rescoring and relabelling models are illustrated in Sections V and VI, including some results and comparisons with the baseline detectors (*i.e.*, Faster RCNN and YOLO).

II. Related Work

A. Object Detection

Interestingly, a leap in the performance of object detection and recognition methods took place from 2012, when Convolutional Neural Networks (CNNs) were reintroduced. Ross *et al.*[21] claim that *"progress has been slow during 2010-2012, with small gains obtained by building ensemble systems and employing minor variants of successful methods"*. CNNs were first proposed in 1998 [27], but they were not widely used due to the limited improvements in computers and datasets. However, since 2010, due to the emergence in computers and datasets, CNNs have been the dominant in the computer vision tasks and the state-of-the-art detectors [46]. Typical CNNs consist of some main layers; input layer, convolutional layer, pooling layer, fully-connect layer and an output layer, as illustrated in Figure 3. In 2012, Alex Krishevsky proposed a CNN architecture interspersing five convolutional layers with max-pooling layers, followed by three fully-connected layers [26]. Noticeable improvement is seen in the field of computer vision since then, hence, CNNs have been widely used and enhanced.

CNNs-based detectors can be categorized into two main groups; *1) Two-Stage Detection* and *2) One-Stage Detection* [30]. The former is known as Region Based CNNs, which includes a prepossessing step where features are extracted and then passed into the pipeline. The latter is called unified pipeline. As its name suggests, this group of CNNs has only one stage where feature extraction, predication of classes are happening. Below, a brief about these groups is presented.

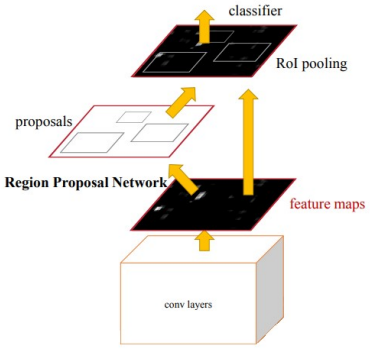


Fig. 4: FasterRCNN Structure, adopted from [42]

Two-stage detectors process detection in a "coarse-to-fine" manner, meaning that they process images in low-resolution then gradually increasing the resolution and propagating the results to the 'finer' image. Detection process occurs in two steps: first, the model proposes a number of regions of interest (RoIs) using some algorithms (e.g. selective search [43]), where those regions are then feed into the CNNs networks. Second, classifiers are applied to provide boundary boxes and probabilities for each detected objects. Some of this group detectors are RCNN [21], Fast RCNN [20], Faster RCNN[42], RFCN [15] and Mask RCNN [23]

- Faster RCNN:

Faster RCNN, proposed by Shaoqing *et al.*[42], uses a network that can be trained to take features and inputted into an ROI pooling layer, hence, it feeds the entire image into the CNN, where regions are extracted and then fed to other layers (ROI pooling, fully connected layers). An illustration of Faster RCNN structure is presented in Figure 4. Faster RCNN has been implemented in a variety of articles studying the importance of contextual information (e.g. [24, 25]), and it is still one of the state-of-the-art detection methods, and due to some of its advantages (e.g. speed, accuracy), it is also used as the baseline detector in this paper.

One-stage detectors as also known as unified pipeline detectors. This group of CNNs has only one stage where feature extraction, predication of classes are happening. YOLO (You Look Only Once) [41], SSD [31] and RetinaNet [29] are examples of this group of detection.

- YOLO:

YOLO was proposed by [41] in 2015. It is considered as the first detector of this detection group. It processes image by dividing it into regions then predicts bounding boxes and probabilities for each region simultaneously, where those processes are taking place in one single network. YOLO is fast, and thus it can be used as real-time detector. It was tested on VOC07 obtaining mAP as 52.7%. Refer to Figure 5 for an illustration presenting YOLOv1 architecture.

For a comparison and discussion of those two groups of detection, we refer the reader to [46, 47].

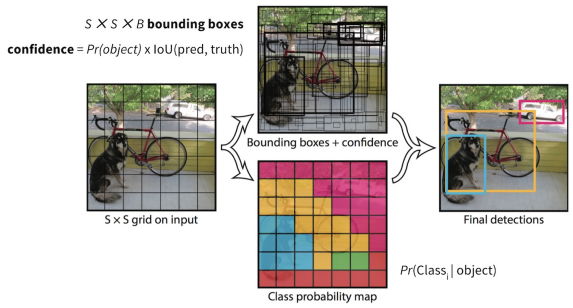


Fig. 5: YOLO Structure, adopted from [41]

B. Contextual Information

Contextual information is defined as *"a statistical property of the world we live in and provides critical information to help us solve perceptual inference tasks faster and more accurately"* [37]. We add that *contextual information is any data obtained from an object's own statistical property and/or from its vicinity including intra-class and inter-class details*. Such a definition is claimed due to the information we observed while studying the importance of context in digital images.

It is said that contextual information is a tool used more with multiple objects, so that relationships among objects can be deeply understood [16]. Roozbeh *et al.*[37] also state that in digital images, objects with clear appearance (e.g. large objects) are easy to detect, whereas some small objects are harder. Lubor *et al.*[39] also claim that contextual information, therefore, can be a solution here as it provides stronger cues in detecting small objects due to the context where those objects are present. Hence, contextual information is described as *"a natural way to improve detection"* [3].

Contextual information in the field of object detection can help to understand and explore object vicinity (*i.e.*, scene-level context) as applied in [4, 45], and also provides object-object relationships (*i.e.*, object-level context) as in [9, 40]. Moreover, Contextual information has been also studied in different areas, such as object localization [12], image segmentation [22], out-of-context detection [11], image annotation [34], scene modeling [6], image understanding [14] and cognitive robotics [48].

- Types of Contextual Information:

Context can be classified upon the sources of information extracted from images. Biederman *et al.*[5] state that there are five categories of object-environment dependencies, which are categorized as *"(i)* ***interposition*** *objects interrupt their background, (ii)* ***support****: objects often rest on surfaces, (iii)* ***probability****: objects tend to be found in some environments but not others, (iv)* ***position****: given an object in a scene, it is often found in some positions but not others, and (v)* ***familiar size****: objects have a limited set of sizes relative to other objects"*. Galleguillos *et al.*[18] grouped those relationships into three main categories, which are: (i) **Semantic** (Probability), (ii) **Spatial** (interposition, support and position) and (iii) **Scale** (familiar size).

III. Proposed Contextual Relationships

In this work, we are following the division of contextual information types proposed by [18]. However, we propose sixteen contextual relationships, as discussed below.

A. Category One: Semantic Context

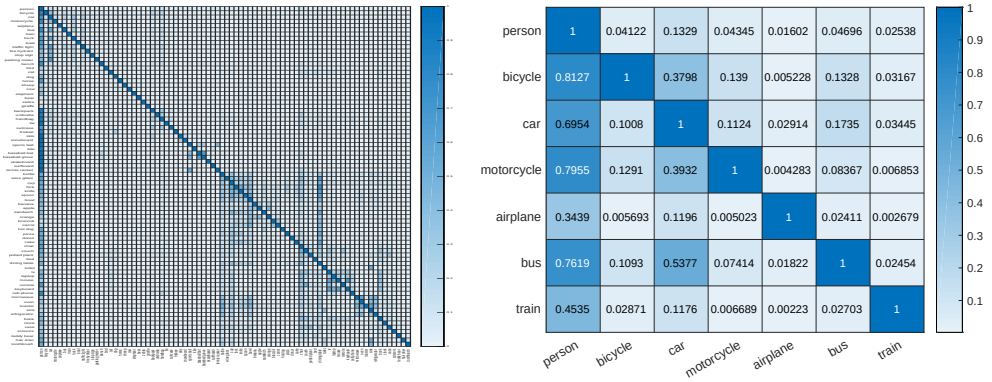
Semantic context, known also as the co-occurrence statistics, records whether objects classes statistically tend to occur in the same scenes. Semantic context is defined as *"the likelihood of an object to be found [presented] in some scenes but not others"* [18]. Such contextual information encodes co-occurrence statistics among objects, thus we can have a clear picture of objects that are more likely to appearing in the same images. This can, therefore, help detectors to refine confidences. Semantic relationship has been widely studied and implemented in a variety of studies, showing an improvement in detection performance such as in [36, 40]. Andrew *et al.*[40] state that semantic contextual information is essential, as it helps minimising the ambiguity in objects' visual appearance, as it was applied as a post-processing tool with local detectors showing that semantic *"greatly improves categorization accuracy"*.

In this work, semantic relationship is the first relationship applied. We analysed the training dataset isolating images with single object (*e.g.*, images with more than one objects are used). A matrix, as shown in Figure 6 is built from the MSCOCO2017 training images following the information presented in the annotations. If object X is presented with object Y, then co-occurrence among objects is positive. If both objects are within the same category (*i.e.*, person) such as a man and a child, co-occurrence will be considered positive as well. This information is used to obtain an overall view of the co-occurrence statistical in MSCOCO dataset, which helps to build the proposed models, as presented in Sections IV and VI.

B. Category Two: Spatial Context

Spatial context is defined as *"the likelihood of finding an object in some position and not others with respect to other objects in the scene"* [18]. Spatial information provides deep information about scenes more than semantic as it concerns not only the co-occurrence (*i.e.*, in an implicitly manner), but also the location and relationships among objects (*e.g.*, a car is above the road).

Although semantic relationship provides a strong cue for disambiguating objects, adding more relations is expected to improve detection even further. Spatial relationships have also been examined and studied in several researches. According to Moshe and Shimon [2], who examined the consequences of pairwise spatial relations between objects, suggesting that encoding proper spatial relations among objects may decrease error rates in recognizing objects. Many studies have included spatial context concerning only *above, below, left* and *right* relationships such as [37]. Others also added other types of spatial features such as *around and inside* [19]. Lu *et al.*[33], furthermore, add more relations such as *taller than, pushing,*



(a) All Objects Co-occurrence Matrix

(b) Seven Objects Co-occurrence Matrix

Fig. 6: Co-occurrence Matrix

carrying. Choi *et al.*[10] propose a contextual relationships model that is developed to leverage co-occurrences and spatial relationships among objects, using a tree graphical model, which is built to encode the dependencies among objects, as parent-child pairwise relationships. The output of this developed model is then combined with the outputs of the local detectors and the global image features. As reported, applying this model increases the performance of detection. In this work, as we believe spatial information is a vital cue in improving detection, we propose four novel sub-groups of spatial information among the reference object and all other detected object. Note that the reference object is chosen by the model to compute its relationships with all other detected objects, where the same process is repeated for all detected objects. Upon the best of our knowledge, none of those relationships have been observed in the literature review.

- Boundary Spatial Relationships:

Boundary relationships consists of four relationships (*i.e.*, *above, below, left, right*). They occur when the boundary of the reference object is not attached or overlapped with other objects. In other words, this contextual constrains represents relationships between objects when there is a gap between objects' boundaries—see Table I for the mathematical equations.

- Central Spatial Relationships:

Similarly, the central spatial relationship calculate relationship upon the centers of the reference object and other objects. For example, central above relationship occurs when the center and the top boundary of the reference object is above the center and the top boundary of the other object, respectively. It can be said that this relationship may include several relationships such as overlapping and size. We say *yes*, and this is what makes it unique as it focuses on the centers of objects but also encodes other relationships that are implicitly counted—see Table I for the mathematical equations used for group of relationships.

- Distance Relationships:

Distance relationship between objects is expected to enrich the context and provide deep knowledge about objects. Distance relationships consist of two types, which are *near* and *far*. The distance is measured to be near/far upon the diagonal of the reference object, as in Table I. If the boundary of X object is far by a distance that is larger than diagonal of the reference object, relationship is considered far, and vice versa.

TABLE I: Spatial Relationships Mathematical Equations.

Boundary Relations	
Above	$(Ref_y + Ref_h) < Obj_y$
Below	$Ref_y > (Obj_y + Obj_h)$
Left	$(Ref_x + Ref_w) < Obj_x$
Right	$Ref_x > (Obj_x + Obj_w)$
Central Relations	
Above	$((Ref_y + Ref_h) \times 0.5) < ((Obj_y + Obj_h) \times 0.5)$ where $Ref_y < Obj_y$
Below	$((Ref_y + Ref_h) \times 0.5) > ((Obj_y + Obj_h) \times 0.5)$ where $(Ref_y + Ref_h) > (Obj_y + Obj_h)$
Left	$((Ref_x + Ref_w) \times 0.5) < ((Obj_x + Obj_w) \times 0.5)$ where $Ref_x < Obj_x$
Right	$((Ref_x + Ref_w) \times 0.5) > ((Obj_x + Obj_w) \times 0.5)$ where $(Ref_x + Ref_w) > (Obj_x + Obj_w)$
Distance	
Near	$(Ref_x - (Obj_x + Obj_w)) < \sqrt{(Ref_w)^2 + (Ref_h)^2)}$
Far	$(Ref_x - (Obj_x + Obj_w)) > \sqrt{(Ref_w)^2 + (Ref_h)^2)}$
Overlapping	
Yes	$Overlapping > 0$
No	$Overlapping < 0$

- Overlapping Relationship:

Overlapping relationship, which is the fifth sup-group of spatial relationships, measures whether the reference object is overlapping with other objects or not, thus, it consists of two types (positive and negative overlapping). Overlapping ratio, as in Equation 1, is considered positive when the Intersection over Union (IoU) value is 0.5 or above.

$$IoU = \frac{area(Ref \bigcap obj)}{area(Ref \bigcup obj)} \tag{1}$$

C. Category Three: Scale Context

Scale contextual information concerns the size of the reference object with respect to other objects in the scene. It has been studied in many researches such as [3, 11, 12, 18, 37]. Those proposed scale relationships are divided into three groups, which are *larger*, *smaller* and *equal*. Refer to Table

TABLE II: Scale Relationships Mathematical Equations.

Scale Context	
Larger	$\sqrt{(Ref_w)^2 + (Ref_h)^2)} > \sqrt{(Obj_w)^2 + (Obj_h)^2)}$
Smaller	$\sqrt{(Ref_w)^2 + (Ref_h)^2)} < \sqrt{(Obj_w)^2 + (Obj_h)^2)}$
Equal	$\sqrt{(Ref_w)^2 + (Ref_h)^2)} = \sqrt{(Obj_w)^2 + (Obj_h)^2)}$

II for equations used to measure scale contextual information among objects. Measurement in this case are counted only upon the diagonals of objects. If the diagonal of reference object is larger than the diagonal of X object, relationship will be considered *larger*. We propose three scale relationships as we expect that the more we study objects context, the more details and knowledge about the scene we obtain.

IV. Proposed Contextual Models Method

Contextual information is an effective property that improves object detection performance. In this section, we are showing how the proposed re-scoring model, which exploit semantic, spatial and scale contexts, improves the detection capacity and analysis various properties of the contextual object detection problem, as experimented on MSCOCO datasets.

Nowadays, as mentioned earlier, CNNs-based detectors have been widely used in the field of object detection, as they are performing as the-state-of-art detectors. However, contextual constrains are still not fully employed by such detectors. They mainly depend on regions of interests, which do not include contextual information.

There are some models proposed to address such an issue by incorporating contextual information into the detection processes by adding further layers into their CNN networks such as [4, 28]. However, as claimed by [3], those models even though seem to improve the detection performance by including context, the contextual knowledge included are *"unclear"*. They are still unable to *"reason about object relations in a manner invariant to viewpoint"*, where they require *"all meaningful relations between all groups of objects [to be observed] from all relevant viewpoints"* in the training data. Therefore, it seems a need to develop a model that can leverage contextual information, where contextual relationships among objects are clearly presented, as we are proposing in this paper.

In the remainder of this section, we detail our rescoring model and analyse its performance. As shown in Figure 7, the model, in the training stage, first takes images from the training dataset, then passes it into the baseline detector. Detector, then, produces the prediction including the bounding boxing and objects labels, which are then encoded and inputted into the classifiers. Encoding features is explained Section IV-A. In addition, in the testing stage, similar steps are followed, but rather than taking an image from training, it is taken from the validation dataset, where the trained classifier applied produces the new scores as the output of the model.

A. Encoding Classifier Inputted Features

Features inputted into the classifiers in both training and testing stages are encoded as follows. Once the detector

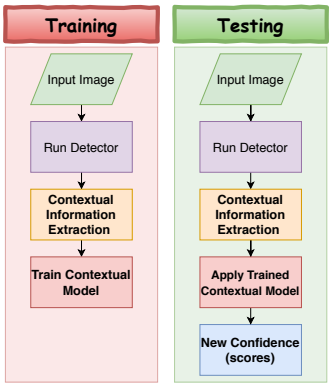


Fig. 7: Procedure of the Proposed Re-scoring Model

TABLE III: Length of Feature Vector Per Relation.

Relationship	Number of features per relation	Length of the feature vector
Co-occurrence	1 (either co-occur or not)	81
Overlapping	2 (Yes, No)	161
Scale	3 (Large, Small, Equal)	241
Spatial 1	4 (above, Below, Left, Right)	321
Spatial 2	4 (above, Below, Left, Right)	321
Near Far	2 (Near, Far)	161
All Relations	**Sum of all above**	**1281**

outputs the bounding boxing and object confidence scores, the desired relationship(s) is/are calculated following the mathematical equations presented in Tables I and II. The length of those features varies upon the number of relationships used. In other words, the length of the feature vector inputted into the classifiers is the confidence value of the reference object + (the length of relationships $\times$ the number of the objects as in MSCOCO annotations (*i.e.*, 80)), as presented in Table III. For example, the length of feature vector in terms of the co-occurrence relationship is 1 + (1 $\times$ 80) = 81, if there is an image containing two objects only, the feature vector is still 81, as only positions of detected objects are encoded and others are zeros.

B. Classifier

For the experiments in this paper, we use a *trainscg* (scaled conjugate gradient back-propagation) Neural Network approach, as implemented in MATLAB [35]. Scaled conjugate gradient (SCG), a supervised learning algorithm, is a network training function used to update weight and bias value according to the scaled conjugate gradient method [38]. *trainscg* was implemented as explained in [35]. The standard network consists of a two-layer feed-forward network, with a sigmoid transfer function in the hidden layer, and a softmax transfer function in the output layer. Several numbers of hidden neurons were tested (*i.e.*, 25, 50, 80, 100, 150, 200, 500, 1,000, 2,000, 5,000), and classifiers with 1,000 Hidden Neurons (HNs) performed the best, as presented in Figure 8. Therefore, 1,000 (HNs) classifier is chosen to be used in all experiments presented in this paper, which is used to re-score detected objects confidence and relabel them.

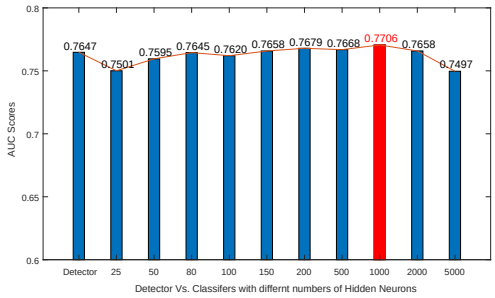


Fig. 8: Implementation of the classifier with different numbers of Hidden Neurons on MSCOCO2017

TABLE IV: AUC Scores: All-Relationship Vs. Faster RCNN.

Threshold Value	Baseline detector	All-relationships model
0.7	0.76472	**0.77057**
0.6	0.77911	**0.78562**
0.5	0.79303	**0.80423**

V. Rescoring

In this section, several experiments have been attempted to examine the impact of the proposed re-scoring model on the Faster RCNN detector using MSCOCO dataset.

A. Contextual Relations Analysis

In this experiment, we examine each relationship and a combination of relationships to investigate their impacts on the performance of the detection, and how they can re-score detected objects' confidences based on context. As presented in Table VII, AUC scores for the relationships and the baseline detectors are presented, where most relationships models over-perform the detector, whereas, we can also see that detector shows better scores in some cases (*e.g.*, Boundary Relations) which could be due to the high variations between the contextual relationships among the detected objects. Standard Deviation (STD) values for each relationship is also presented as shown between brackets to show the difference in scores where five trials were used for each relationship.

B. Combined model

As shown in Section V-A, the proposed re-scoring models obtain higher AUC scores than the baseline detector in the majority of the cases. We, therefore, combined all relationships into one model. Detector threshold values are set as [0.5, 0.6, 0.7] in this experiment, because we assume applying different threshold values may enable the detector, in some cases, to detect more objects. Table IV shows a comparison in AUC scores between our approach (all-relationships model) and Faster RCNN detector. AUC scores of our contextual model is higher than the detector in all three cases. Figure 9, furthermore, shows some outputs for our model to illustrate the performance compared with the detector on how objects confidences are re-rated based on their vicinity.

Noticeably our model drops the the scores of the `dining table`) from 0.7143 to 0.2166, which is incorrectly detected.

TABLE V: AUC Scores: Re-scoring Model Vs. YOLOv1

Baseline detector (YOLOv1)	All-relationships model
0.66977	**0.67894**

TABLE VI: AUC Scores: Re-scoring Model Vs Faster RCNN on PASCAL Dataset

Baseline detector (Faster RCNN)	All-relationships model
0.78432	**0.79369**

We assume `dining table` often appears in different spatial and scale configuration with regards to other detected objects.

YOLOv1 is also examined, as it is a detector that implicitly includes contextual information into its end-to-end detection pipeline. The proposed re-scoring model performs better than this baseline detector using MSCOCO2017 dataset, as presented in Table V, where only one threshold value (*i.e.*, value is 0.7) is used.

Furthermore, we also examined our re-scoring model on PASCAL2012 dataset [17] where the baseline detector is Faster RCNN (threshold value is 0.7). As presented in Table VI, our model also performs better than Faster RCNN on this dataset.

C. Out-Of-Context

As preset ed earlier, all-relationship model (*i.e.*, re-scoring model) shows good results, and outperforms the performances of Faster RCNN on both MSCOCO2017 and PASCAL2012, and YOLOv1 on MSOCOC2017. Therefore, we deeply studied the model to trying to examine if the performance would be even higher by iterating the process of re-scoring, where this experiment is only run on Faster RCNN as the baseline detector. Yes, the performance has increased from 0.77057 to 0.80702 (in the 12^{th} iteration), which is a great step in re-scoring the detection outputs, as presented in Figure 10, where

TABLE VII: AUC Scores and STD: Different Relationships

One Contextual Relationship Model	
Relation	AUC Scores (STD)
Co-occurrence	0.766 (0.0015)
Boundary	0.758 (0.0016)
Central	0.758 (0.0011)
Overlapping	0.773 (0.0017)
Near/Far	0.766 (0.0010)
Scale	0.766 (0.0012)
Detector	0.764

Two Contextual Relationships Model					
Relations	Boundary	Central	Overlapping	Near/Far	Scale
Co-occurrence	0.763 (0.0010)	0.764 (0.0015)	0.772 (0.0012)	0.767 (0.0017)	0.758 (0.0017)
Boundary	-	0.756 (0.0005)	0.771 (0.0011)	0.767 (0.0018)	0.768 (0.0013)
Central	0.756 (0.0005)	-	0.767 (0.0010)	0.752 (0.0010)	0.766 (0.0017)

Three Contextual Relationships Model	
Relations	AUC Scores (STD)
Co-occurrence +Boundary+Scale	0.768 (0.0018)
Co-occurrence +Central+Scale	0.765 (0.0015)
Co-occurrence +Boundary+Central	0.759 (0.0019)
Boundary+Central+Scale	0.768 (0.0021)

Four Contextual Relationships Model	
Relations	AUC Scores (STD)
Co-occurrence +Boundary+Scale+Overlapping	0.771 (0.0007)
Co-occurrence +Central+Scale+Overlapping	0.767 (0.0017)

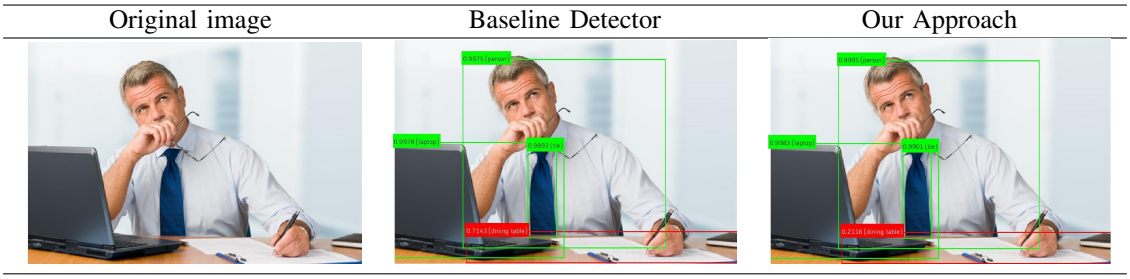


Fig. 9: All-Relationships model vs. Detector outputs: green boxes represent correct detection, whereas red are incorrect (not in the ground-truth)

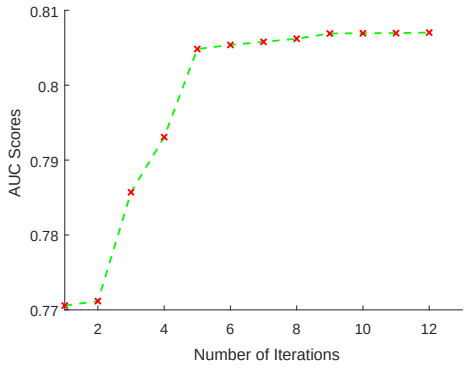


Fig. 10: Running the Re-scoring Model in Iterations

the AUC scores increase from the 1st run to 12th iteration run. This iterated process is done as follows: First, we apply the rescoring model (*i.e.*, all-relationship model) to the detector outputs. Scores obtained from the rescoring model are then fed again to the rescoring model for 12 times, where the final outputs (*i.e.*, the 12th run outputs) are considered as the iterated model outputs.

Furthermore, Figure 11 shows results of running the model in 12th iterations, where the detector output is also shown. As presented, the scores change in the 1st run, but even improved in the 12th iteration run. *Zebra* objects scores have increased, where one instance reaches 1, which is likely to occur due to the context presented and the other detected objects present. However, *elephant* class score was dropped considerably as in the 1st run, and even more in the 12th iteration, where this object was detected incorrectly.

Due to the success obtained during running the model in iterations, examining whether this model can be an effective tool in re-rating out-of-context objects. The answer, as shown in Figure 12, is **yes**. In the first row, we can see how the model decreases the confidences of *cat* due to its location and scale. However, it can be seen how cars and persons confidences increased apart from cars that overlap with the cat (e.g. the car next to the cat head), which were lower compared to other cars. We assume this is because the detector is 2D based and having a car in such location compared with the cat is unlikely.[1] This is a great result that the model shows, which can be said that this model is a good tool for out-of-context objects. In the second row, the model reduces the confidence of *elephant*, which we assume is due to the present of *couch*. This may raise a question, why only the elephant score was reduced, and not the couch, as person usually appears with both objects. We suggest that the answer is that due to the location and the size of person compared with the couch, the elephant is seen out-of-context in this scene, and this could be why the model reduces its confidence.[2]

VI. Re-labelling

In this experiment, we researched further on how to improve the performance, and upon the success of the proposed contextual re-scoring model (Section V), we decided to move one step further to not only rescore objects confidences but also to re-label them upon their contextual vicinity.

A. Relabelling Model

In this experiment, Faster RCNN is used as the baseline detector and examined on MSCOCO2017. This is implemented as follows. *First*, we set a minimum threshold value for our contextual re-rating model as 0.4. *Second*, any detected objects re-scored by the re-scoring model with less than the threshold are passed into our relabelling model. *Third*, the top five possibilities obtained from the detector including the reference objects are passed into our re-scoring model. If any of the possibilities are re-scored with a higher value than threshold, then the object(s) with maximum value is considered as the new labelled object, if none is higher, then the reference object will be removed and considered as background. *Fourth*, after new labels are determined, all objects including the new labels are passed again into our re-scoring model, to obtain the new confidences.

The proposed relabelling model is illustrated in Figure 13, where the process from inputting the images until outputted are shown. Note that all steps in the red squared are the core processes involved in this approach.

Furthermore, re-labelling model, as presented in Table VIII, obtains higher AUC scores than the baseline detector and the re-scoring model. This is because the proposed re-labelling model is not only re-rating objects confidences, but also suggesting new objects labels and removing objects with lower confidences than the set threshold value, based on the

[1]Image is taken from Instagram with permission from account @fransditaa

[2]Image is taken from SUN dataset: Out-of-context images [44].

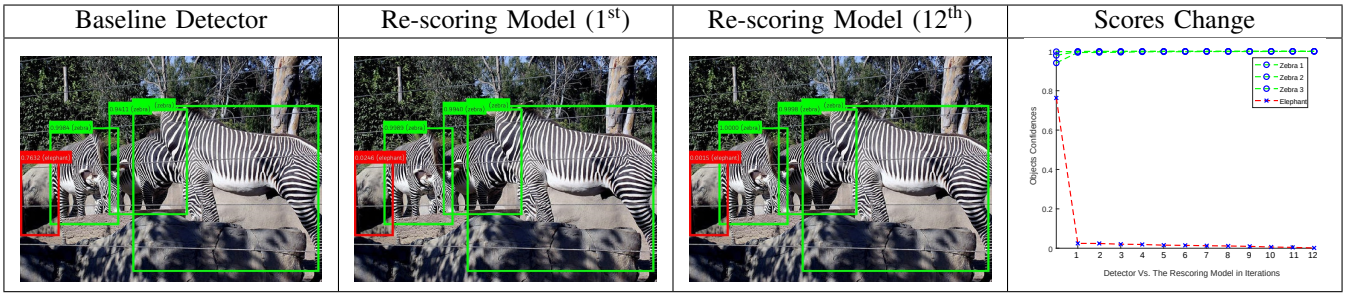


Fig. 11: Results: Running the re-scoring model in Iterations

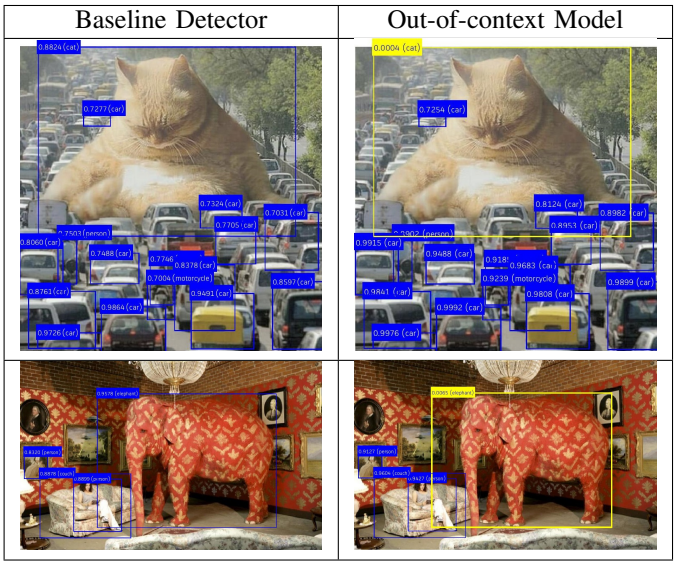


Fig. 12: Results: Out-of-Context: Yellow boxes represent out-of-context objects, blue boxes represent other objects regardless whether they are correctly detected or not.

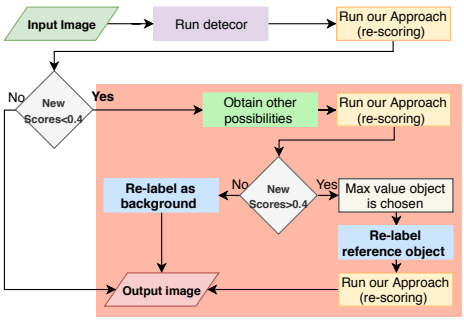


Fig. 13: Relabelling Approach

contextual information encoded from the scene. In addition, we also use average precision (AP) and its mean (mAP), where IoU threshold is 0.5, and F1 score as other evaluation metrics to show the effectiveness of this proposed re-labelling model. Results of using such evaluation metrics are presented in Table IX. The relabelling model achieves a better performance than the baseline detector in terms of improving both mean average precision (mAP) and Fl.

Table 14 shows some results of the proposed relabelling

TABLE VIII: AUC Scores: Faster RCNN VS. Re-Scoring and Relabelling.

Threshold Value	Detector	Re-Scoring	Relabelling
0.7	0.76472	0.77057	**0.78278**
0.6	0.77911	0.78562	**0.79446**
0.5	0.79303	0.80423	**0.81084**

TABLE IX: AP and F1 scores in percentages [%] for the baseline detector and our proposed re-labelling model.

Threshold Value	**Baseline Detector**		**Re-labelling Model**	
	$mAP_{0.5}$	F1	$mAP_{0.5}$	F1
0.7	62.82	57.34	**65.50**	**58.95**
0.6	57.55	52.77	**64.14**	**56.35**
0.5	51.38	48.68	**63.14**	**55.02**

model outputs. Starting from the first row, where threshold value for the detector is 0.6. We can see how the relabelling model corrects labelling *chair* to *bench*. In details, the detector detects the two cats correctly which confidences of 0.9955 and 0.9964, but incorrectly detects the bench and as a chair with a confidence of 0.6294. The re-scoring model is run, which decreases the chair scores to 0.0816, making it ready for the relabelling model to figure out if there are another object that is more likely to fit. The relabelling model is run to improve the detection, which then suggests the bench with confidence of 0.8485.

Detector with threshold of 0.7, as the main baseline detector of all experiments in this work, is run as well. As seen in the second row, where the model correctly relabels three incorrectly detected objects (*i.e.*, 2 carrots and a book). A knife was incorrectly detected as a book, where the re-scoring model re-rates as 0.2149, then the relabelling model relabels as a knife with confidence of 0.7895. Similarly, in the third row. The relabelling model removes the incorrectly detected objects (*i.e.*, traffic light), where it also increases the confidences of the correct detected objects. We believe this great improvement is achieved by the model due to the spatial and scale relationships, particularly the overlapping relationships among the detected objects. In other words, traffic lights are unlikely to be smaller than person and overlapped in such a way. Because the detector did not suggest a better fitting for such objects in such location, the model suggests relabelling as background, which results to correct detection.

TABLE X: AUC scores for some MSCOCO object classes: Faster RCNN and Re-Scoring and Relabelling Models.

Class ID	Class Label	Detector	Re-Scoring	Relabelling
1	Person	0.84345	**0.84646**	0.84568
2	Bicycle	0.76073	0.76984	**0.78006**
10	Traffic light	0.78979	0.78690	**0.79083**
18	Horse	**0.92223**	0.89649	0.88535
:	:	:	:	:
34	Kite	0.73866	0.76852	**0.78292**
:	:	:	:	:
43	Fork	0.66847	0.79619	**0.81545**
44	Knife	**0.69595**	0.59494	0.56944
:	:	:	:	:
72	Sink	0.76167	0.78766	**0.79756**
78	Teddy bear	**0.80049**	0.78015	0.77897
Mean	-	0.76472	0.77057	**0.78278**

It is usually said that perfection is unattainable, which is the case in applying to this relabelling model. Yes, the relabelling model still makes mistakes. As illustrated in the last row, we can see that the relabelling model suggests a kite instead of surfboard, making this incorrect detection, which also reduces all objects scores including the correctly detected objects.

Furthermore, due to the high number of object classes in MSCOCO, only a few objects AUC scores are presented for a comparison between baseline detector, re-scoring model and re-labelling model as in Table X, where detector threshold value is 0.7.

B. Iterated Relabelling Model

Upon the success that the relabelling model shows (Section VI-A), it is expected to obtain a much higher performance when applying the relabelling model on the top of the iterated rescoring model due to the great performance it shows (V-C), compared to the re-scoring model with no iteration (V-B). As presented in Figure 10, the re-scoring model performance increases when iterated, until the 12^{th} iteration, due to this great performance, applying the relabelling on the 12^{th} iterated re-scoring model, named *Iterated Relabelling Model (IRM)* is expected to perform greater than when applied on the plain re-scoring model. Table XI shows the AUC scores comparing the performance of the baseline detector (Faster RCNN) vs. the iterated re-labelling model. It can be seen that the IRM shows a very great increased performance compared to the detector, which is expected, as it corrects labels and changes object confidences.

In addition, as the IRM has corrected object labels, it seems essential to present the F1 and $mAP_{0.5}$ and mAP (IoU=[0.5:0.05:0.95]) scores. Table XII shown the F1 and mAP scores in percentage for the baseline detector, relabelling model and iterated relabelling model, where the threshold used for the detector is 0.7. IRM obtains the highest scores with an improvement of 8% is achieved compared to the detector performance.

Figure 15 shows some results obtained when IRM is applied. As the comparison in this figure is intended only to compare the outputs of the detector (*i.e.*, Faster RCNN) and IRM, only their outputs are illustrated. The results obtained from Faster RCNN and IRM are shown in the left and right

TABLE XI: A comparison between 10 MSCOCO object classes AUC scores for Faster RCNN (threshold is 0.7) and the iterated Relabelling Model.

Class ID	Class Label	Faster RCNN	IRM
1	Person	0.84345	**0.99955**
2	Bicycle	0.76073	**0.99170**
:	:	:	:
5	Airplane	0.88300	**1**
21	Elephant	0.74058	**0.96282**
25	Backpack	0.61494	**1**
27	Handbag	0.67471	**1**
:	:	:	:
37	Skateboard	0.84024	**0.97560**
44	Knife	0.69595	**0.97878**
45	Spoon	**0.65490**	0.57254
:	:	:	:
67	Keyboard	0.75925	**0.96692**
Mean	-	0.76472	**0.95314**

TABLE XII: AP and F1 scores in percentages [%] for the baseline detector and our proposed re-labelling models.

Model	$mAP_{0.5}$	mAP	F1
Baseline Detector	62.82	33.48	57.34
Re-labelling Model	65.50	34.07	58.95
Iterated Re-labelling Model	**70.10**	**40.43**	**64.84**

columns respectively. In the first row, three objects are detected (*i.e.*, a person, a sports ball, and a tennis rackets). All objects were correctly detected with confidence higher than 0.94. Once inputted into the IRM, the model even increases the performance of each object with a minimum of 0.98. This can be due to the high semantic, spatial and scale relationships. They are likely to appear in real-world in such position and scale. In the second row, 5 objects were detected, 3 are correctly detected (*i.e.*, a person and 2 elephants) and the other 2 are incorrect (*i.e.*, a person and a cow). Even though, Faster RCNN detector detects all objects with high confidence including the incorrect objects (*i.e.*, minimum confidence is 0.74 for the cow), the IRM discards all incorrect objects and label them as background, whereas, increases the correctly detected objects confidence. This is due to the spatial and scale relationships between person and elephant.

VII. Conclusion

In this paper, we present a machine learning approach that can be integrated into most of object detection methods as a post-processing step, to improve detection performance and help to correct false detection based on the contextual information encoded from the scene. It can also help in lowering out-of-context objects. As illustrated, experimental results show that our models obtain higher AUC scores ($\approx$ 1.8%) compared to the state-of-the-art baseline detectors, as well as higher mAP and F1 scores. This paper shows that semantic, spatial and scale relationships enhance the detection performance, where correcting and relabeling false detection can be also attempted. A deeper investigation of spatial and scale contexts, and the interaction between objects appearances and contextual features are to be explored and modelled as an

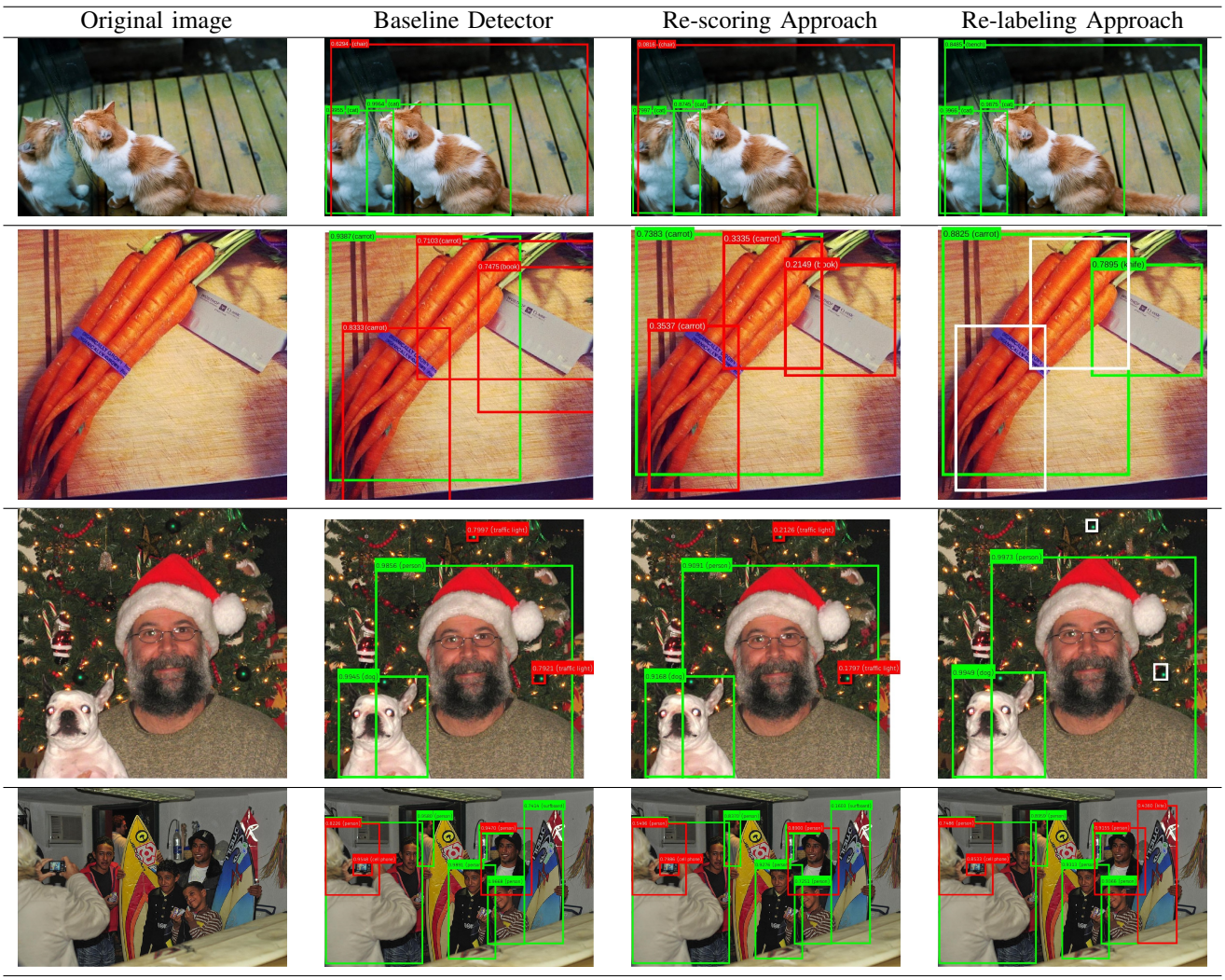


Fig. 14: Threshold 0.5: Relabeling and Re-scoring models outputs: Green, red and white boxes represent correct detection, incorrect detection, and objects removed and re-labelled as background, respectively

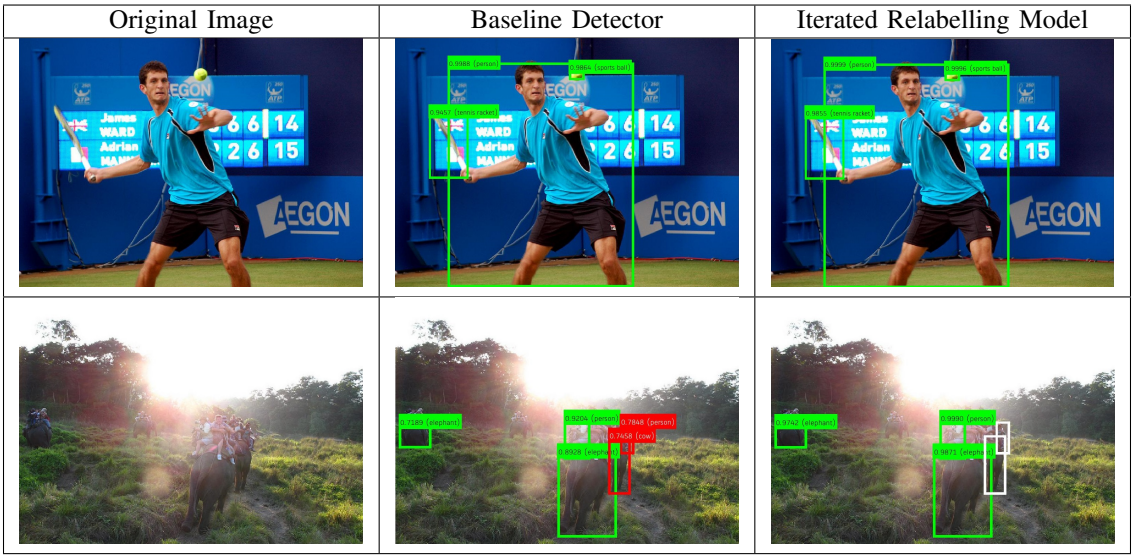


Fig. 15: Results: IRM Results vs Detector Outputs: Green, red and white boxes represent correct detection, incorrect detection, and objects removed and re-labelled as background, respectively

end-to-end pipeline including bounding boxes refinement is preliminary and left as future work.

REFERENCES

[1] Faisal Alamri and Nicolas Pugeault. Contextual relabelling of detected objects. *CoRR*, abs/1906.02534, 2019.

[2] Moshe Bar and Shimon Ullman. Spatial context in recognition. *Perception*, 25(3):343–352, 1996. PMID: 8804097.

[3] Ehud Barnea and Ohad Ben-Shahar. Contextual object detection with a few relevant neighbors. *CoRR*, abs/1711.05705, 2017.

[4] Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross B. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. *CoRR*, abs/1512.04143, 2015.

[5] Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143 – 177, 1982.

[6] Ilker Bozcan and Sinan Kalkan. COSMO: contextualized scene modeling with boltzmann machines. *CoRR*, abs/1807.00511, 2018.

[7] X. Cao, X. Wei, Y. Han, and X. Chen. An object-level high-order contextual descriptor based on semantic, spatial, and scale cues. *IEEE Transactions on Cybernetics*, 45(7):1327–1339, July 2015.

[8] G. Chen, Y. Ding, J. Xiao, and T. X. Han. Detection evolution with multi-order contextual co-occurrence. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1798–1805, June 2013.

[9] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. *CoRR*, abs/1704.04224, 2017.

[10] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 129–136, June 2010.

[11] Myung Jin Choi, Antonio Torralba, and Alan S. Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853 – 862, 2012. Special Issue on Awards from ICPR 2010.

[12] Myung Jin Choi, Antonio Torralba, and Alan S. Willsky. A tree-based context model for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):240–252, Feb 2012.

[13] Ramazan Gokberk Cinbis and Stan Sclaroff. Contextual object detection using set-based classification. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 43–57, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[14] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. *CoRR*, abs/1704.03114, 2017.

[15] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409, 2016.

[16] Chaitanya Desai, Deva Ramanan, and Charless C. Fowlkes. Discriminative models for multi-class object layout. *International Journal of Computer Vision*, 95(1):1–12, Oct 2011.

[17] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, Jun 2010.

[18] Carolina Galleguillos and Serge Belongie. Context based object categorization: A critical survey. *Comput. Vis. Image Underst.*, 114(6):712–722, June 2010.

[19] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. Object categorization using co-occurrence, location and appearance. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[20] Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015.

[21] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[22] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80(3):300–316, Dec 2008.

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[24] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. *CoRR*, abs/1711.11575, 2017.

[25] N. Inoue, R. Furuta, T. Yamasaki, and K. Aizawa. Object detection refinement using markov random field based pruning and learning based rescoring. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1652–1656, March 2017.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[28] Jianan Li, Yunchao Wei, Xiaodan Liang, Jian Dong, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Attentive contexts for object detection. *CoRR*, abs/1603.07415, 2016.

[29] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

[30] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul W. Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *CoRR*,

abs/1809.02165, 2018.

[31] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[32] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. *CoRR*, abs/1807.00119, 2018.

[33] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016.

[34] Zhiwu Lu, Horace H. S. Ip, and Yuxin Peng. Contextual kernel and spectral methods for learning the semantics of images. *IEEE Transactions on Image Processing*, 20(6):1739–1750, June 2011.

[35] MATLAB. *Deep Learning Toolbox R2017a*. The MathWorks Inc., Natick, Massachusetts, United States, 2017.

[36] T. Mensink, E. Gavves, and C. G. M. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2441–2448, June 2014.

[37] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, June 2014.

[38] Martin Fodslette Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525 – 533, 1993.

[39] Dennis Park, Deva Ramanan, and Charless Fowlkes. Multiresolution models for object detection. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 241–254, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[40] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge J. Belongie. Objects in context. *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[41] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[42] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[43] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *2011 International Conference on Computer Vision*, pages 1879–1886, Nov 2011.

[44] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, June 2010.

[45] Xingyu Zeng, Wanli Ouyang, Bin Yang, Junjie Yan, and Xiaogang Wang. Gated bi-directional cnn for object detection. In *ECCV*, 2016.

[46] Wang Zhiqiang and Liu Jun. A review of object detection based on convolutional neural network. In *2017 36th Chinese Control Conference (CCC)*, pages 11104–11109, July 2017.

[47] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *CoRR*, abs/1905.05055, 2019.

[48] Hande Çelikkanat, Güner Orhan, Nicolas Pugeault, Frank Guerin, Erol Şahin, and Sinan Kalkan. Learning context on a humanoid robot using incremental latent dirichlet allocation. *IEEE Transactions on Cognitive and Developmental Systems*, 8(1):42–59, March 2016.