Computer Vision and Image Understanding 152 (2016) 118–130

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Recovering hard-to-find object instances by sampling context-based object proposals

José Oramas M.*, Tinne Tuytelaars

KU Leuven, ESAT-PSI, iMinds, Kasteelpark Arenberg 10 - bus 2441, B-3001 Heverlee, Belgium

ARTICLE INFO

Article history:
Received 6 October 2015
Revised 16 August 2016
Accepted 17 August 2016
Available online 18 August 2016

Keywords:
Object detection
Object proposal generation
Context-based reasoning
Relational learning

ABSTRACT

In this paper we focus on improving object detection performance in terms of recall. We propose a post-detection stage during which we explore the image with the objective of recovering missed detections. This exploration is performed by sampling object proposals in the image. We analyse four different strategies to perform this sampling, giving special attention to strategies that exploit spatial relations between objects. In addition, we propose a novel method to discover higher-order relations between groups of objects. Experiments on the challenging KITTI dataset show that our proposed relations-based proposal generation strategies can help improving recall at the cost of a relatively low amount of object proposals.

1. Introduction

Object detection methods have become very effective at localizing object instances in images. Different methods have been proposed, ranging from methods that model the appearance of the object as it is projected on the 2D image space (Dalal and Triggs, 2005; Felzenszwalb et al., 2010; Girshick et al., 2014; He and Sun, 2014) to methods that reason about physical properties of the objects in the 3D scene (Pepik et al., 2012; Zia et al., 2014). All these methods have one thing in common: they rely completely on appearance features, e.g. color, shape or texture, to describe the objects of interest. If the object is clearly visible in the image, appearance cues can be very strong. Unfortunately, appearance-based approaches cannot cope well with more difficult cases, such as small object instances or highly occluded ones. In spite of some efforts in this direction (e.g. Frankenclassifier (Mathias et al., 2013), Occlusion Patterns (Pepik et al., 2013), Occlusion Boundaries (Hoiem et al., 2011)), these mostly remain undetected, resulting in reduced recall. In a real world setting, highly cluttered scenes and therefore small and occluded objects are actually quite common - probably more common than in typical benchmark datasets which are often object-focused (e.g. because they have been collected by searching images that have the object name mentioned in the tags).

In recent years, several works (Choi et al., 2010; Desai et al., 2011; Hoiem et al., 2006; Oramas et al., 2014; Perko and Leonardis, 2010) have proposed the use of contextual information. These works typically follow a two-stage pipeline during testing. First, a set of detections is collected using an appearance-based detector. Then, using a pre-learned context model, out-of-context detections are filtered-out. This strategy has been effective at improving object detection, specifically, in terms of precision. On the downside, objects missed by the object detector are not recovered, which leaves no room for improvement in terms of recall. A possible explanation for this tendency, is that the high-precision low-recall area is often considered the more interesting part of the precision-recall curve (Atanasoaei et al., 2010; Li and Chen, 2010). Methods are optimized and typically perform well in this region. The high-recall low-precision area, on the other hand, receives little attention – as if we all have come to accept there is some percentage of object instances that are just too hard to be found. A very different view is common in the work on class-independent object proposals detection (e.g. objectness (Alexe et al., 2010; 2012a), selective search (Uijlings et al., 2013), edge boxes (Zitnick and Dollár, 2014), deepProposals (Ghodrati et al., 2015) or deepBoxes (Kuo et al., 2015)). When the object class is unknown, no-one expects a high precision, and it is only natural to focus on recall instead. A common evaluation protocol in this context is the obtained recall as a function of the number of window proposals per image. Here, we adopt the same evaluation scheme, but now for standard supervised object detection.

In a sense, this work is similar to Vedaldi et al. (2009) which also focuses on recall instead of precision. The goal is to find as many object instances as possible, even if this comes at a cost, in the form of many false positives (low precision). Because of the lower precision, we refer to the detections as 'object proposals', as in the class-independent object detection work. This

* Corresponding author.
E-mail address: jose.oramas@esat.kuleuven.be (J. Oramas M.).

http://dx.doi.org/10.1016/j.cviu.2016.08.007
1077-3142/

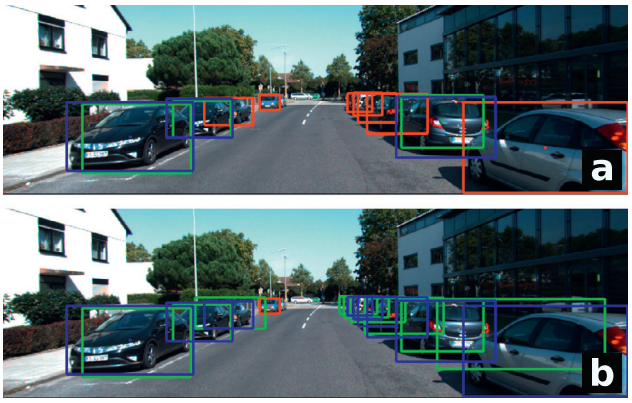

Fig. 1. Object detections collected: a) after running a standard appearance-based detector, b) after sampling 100 context-based object proposals post detection. Notice how we manage to recover many of the initially missed detections. Matched annotations are marked in blue, missed detections in red and matching object proposals in green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

reflects the idea that further verification (e.g. using other modalities, other viewpoints or higher resolution imagery) may be required to separate the true positives from the many false positives – a process which may be application dependent and is out-of-scope of this work. We compare various strategies to generate object proposals: i) a sliding-window baseline, ii) two methods for class-independent object proposals (selective search (Uijlings et al., 2013) and edge boxes (Zitnick and Dollár, 2014)), and iii) two class-specific context-based schemes. Different from Vedaldi et al. (2009), which exploits intrinsic appearance features, we focus on context cues from other objects in the scene. Multiple objects in a scene often appear in particular spatial configurations. This means that detecting one object also provides information about possible locations of other objects. We start from a few high-confidence appearance-based detections and use these as seeds based on which other likely object locations are identified. We explore one method that uses pairwise relations, and propose a new topic-based method that builds on higher-order spatial relations between groups of objects. We have found that, based on very simple features, relative location and pose, our method is able to discover arrangements between objects that resemble those found in the real world. Furthermore, it does not enforce restrictions on the number of objects participating in each of the higher-order relations. For simplicity, we assume the ground plane to be known, both for the baselines as for the newly proposed context-based schemes. However, note that if needed, these can be estimated in different ways (e.g. Bao et al. (2011); Hoiem et al. (2006)). We show that our method is able to bring significant improvement to standard object detectors. For example, notice how in Fig. 1b we manage to recover many of the initially missed detections (Fig. 1a). This is achieved at a relatively low cost of just 100 additional object proposals.

The remainder of this paper is organized as follows: Section 2 presents related work. In Section 3 we present the details of the analysis and of the methods for generating object proposals. Experiments, results and discussions are presented in Section 4. Then, Section 6 addresses the current limitations of the proposed method. Finally, Section 7 concludes this paper.

2. Related work

The analysis presented in this paper lies at the intersection of class-independent and context-based class-specific object detection. These two groups of work constitute the axes along which we position our work.

2.1. Context-based class-specific object detection

Contextual information, in the form of relations between objects, has been successfully exploited to improve object detection performance in terms of precision (Choi et al., 2010; Desai et al., 2011; Felzenszwalb et al., 2010; Perko and Leonardis, 2010). However, objects missed by the object detector are not recovered. This, in consequence, leaves no room for improvement in terms of recall. One work that tries to increase recall is the co-detection work of Bao et al. (2012). They exploit detections of the same object instances in multiple images to generate bounding boxes. Our work, on the other hand, operates on a single image. Different from Mottaghi et al. (2014), our method does not require an image segmentation step. Furthermore, our contextual models are defined in 3D space. This last aspect also separates our method from Li et al. (2014). This makes our models easier to transfer to other datasets with other camera viewpoints and gives them some level of interpretability which can be exploited in other applications, e.g. autonomous driving or robotic manupulation. Different from Yao and Li (2010), which models human-object interactions, our context models are more flexible since the majority of the related objects may not occur in the scene, whereas in Yao and Li (2010), the objects and the parts of the body are always present. Additionally, our work differs from Choi et al. (2010); Desai et al. (2011); Felzenszwalb et al. (2010); Perko and Leonardis (2010) in that we consider higher-order relations whereas most of the methods that exploit relations between objects focus on the pairwise case. Recently, a small group of works (Cao et al., 2015; Oramas et al., 2014; Zhang et al., 2013) that consider higher order relations has been proposed. In Cao et al. (2015), a Pure-Dependency (Hou et al., 2011) framework is used to link groups of objects. In Oramas et al. (2014), objects are grouped by clustering pairwise relations between them. The work of Zhang et al. (2013) is able to reason about higher-order semantics in the form of traffic patterns. Different from these works, our topic-based method to discover higher-order relations does not require the number of participating objects to be predefined (Hou et al., 2011). Furthermore, objects do not need to be "near" in the space defined by pairwise relations in order to be covered by the same higher-order relation (Oramas et al., 2014). Finally, our method does not require scene-specific cues (e.g. lane presence, lane width or intersection type), or motion information (Zhang et al., 2013).

Another related work is Küttel et al. (2010) where two methods are proposed to learn spatio-temporal rules of moving agents from video sequences. This is done with the goal of learning temporal dependencies between activities and allows interpretations on the observed scene. Our method is similar to Küttel et al. (2010) in that both methods perform spatial reasoning and both methods are evaluated in a street scene setting. Different from Küttel et al. (2010) which aims at building scene-specific models, the models produced by our method are specific to the object classes of interest and not scene-dependent. Furthermore, while Küttel et al. (2010) focuses more on motion (flow) cues, our method focuses on instance-based features (location & pose). Moreover, the method from Küttel et al. (2010) requires video sequences and operates in the 2D image space while our method runs on still images and operates in the 3D space.

2.2. Class-independent object detection

Another group of work operates under the assumption that there are regions in the image that are more likely to contain objects than others. Based on this assumption, the problem is then

to design an algorithm to find these regions. Following this idea, Alexe et al. (2010) proposed a method where windows were randomly sampled over the image. Following the sampling, a "general" classifier was applied to each of the windows. This classifier relied on simple features such as appearance difference w.r.t. the surrounding or having a closed contour and was used to measure the objectness of a window. In Alexe et al. (2010), windows with high objectness are considered to be more likely to host objects. Later, Endres and Hoiem (2014) proposed a similar method with the difference that their method generated object proposals from an initial segmentation step. This produced better aligned object proposals. Similarly, Uijlings et al. (2013) proposed a selective search method which exploits the image structure, in terms of segments, to guide the sampling process. In addition, their method imposes diversity by considering segment grouping criteria and color spaces with complementary properties. Recently, Zitnick and Dollár (2014) proposed a novel objectness measure, where the likelihood of a window to contain an object is proportional to the number of contours fully enclosed by it. A common feature of this group of work is that their precision is less critical. The number of generated proposals is anyway only a small percentage of the windows considered by traditional sliding window approaches. On the contrary, these methods focus on improving detection recall by guiding the order in which windows are evaluated by later class-specific processes. In Alexe et al. (2012b), these ideas were integrated in a context-based detection setting where new proposals are generated sequentially based on previously observed proposals following a class-specific context model. Inspired by these methods we propose to complement a traditional object detector with an object proposal generation step. The objective of this additional step is to improve detection recall even at the cost of more false positives. Different from Alexe et al. (2012b), which just returns a single window per image (thus detecting a single object instance), we generate several windows per image with the objective of recovering as many object instances as possible. Moreover, our context information is object-centered.

Recently, Long et al. (2014) proposed "location relaxation", a two-stage detection strategy where candidate regions of the image are identified using coarse object proposals generated from bottom-up segmentations. Then, based on these proposals, a top-down supervised search is performed to precisely localize object instances. Similar to Long et al. (2014) we propose a two-stage strategy to improve object detection. However, instead of focusing on refining object localization our focus is on maximizing the number of detected instances. In this aspect, the proposed method and the work from Long et al. (2014) complement each other since the proposed method can be used to coarsely localize object instances while the strategy from Long et al. (2014) can be used to improve localization accuracy. Another difference, is that while the method from Long et al. (2014) uses local class-specific models to improve the localization of a specific object instance, our method uses context models to explore candidate locations of other instances.

3. Proposed method

The proposed method can be summarized in 2 steps: In a first stage, we run a traditional object detector which produces a set of object detections. Then, in a second stage, we sample a set of object proposals aiming to recover object instances possibly missed during the first stage.

3.1. Class-specific object detection

The main goal of this work is to recover missed object instances after the initial detection stage has taken place. Given this focus on the post-detection stage, for the object detection stage we start from an off-the-shelf detector. In practice, given a viewpoint-aware object detector, i.e. a detector that predicts the bounding box and viewpoint of object instances, we collect a set of 2D object detections $o = \{o_1, o_2, \ldots, o_n\}$ where each object detection $o_i = (b_i, \alpha_i, s_i)$ is defined by its detection score s_i, its predicted viewpoint α_i and its 2D bounding box coordinates $b_i = (x_{1i}, y_{1i}, x_{2i}, y_{2i})$.

3.2. Object proposal generation methods

Traditional appearance-based object detectors have proven to be effective to detect objects o with high confidence for the cases when objects of interest are clearly visible. At the same time, for small or highly-occluded object instances its predictions are less reliable resulting in a significant number of object instances being missed. To overcome this weakness we propose, as a post-detection step, to sample (class-specific) object proposals o' with the goal of recovering missed detections. We analyze four strategies to generate these proposals, as discussed in the next four sections.

Relaxed score detector

A first, rather straightforward method to recover missed detections consists of further reducing the threshold τ used as cutoff in the object detector. This is a widely used strategy, even though it usually does not increase recall that much. We refer to this strategy as *Relaxed Score Detector*. This strategy consists of the original object detector with non-maximum suppression (NMS) performed with default settings while reducing drastically the threshold τ for the detection score.

Relaxed NMS detector

In addition, we define an alternative strategy to relax the object detector. Instead of lowering the threshold τ, we remove the non-maximum suppression step present in most object detectors (including the one used in our experiments). For a given threshold value, this results in many more object proposals being generated. This allows to detect objects highly-occluded by other objects of the same class. We refer to this strategy as *Relaxed NMS Detector*.

3D Sliding window

This is a 3D counterpart of the 2D sliding window approach used by traditional detectors (e.g. Felzenszwalb et al. (2010)). This approach is inspired by the work of Hoiem et al. (2006). We assume the existence of a ground plane that supports the objects of interest. Given the ground plane, we densely generate a set of 3D object proposals $O' = \{O'_1, \ldots, O'_m\}$ resting on it for each of the discrete orientations $\theta_k = \{\theta_1, \ldots, \theta_K\}$. Each 3D object proposal, $O' = (X, Y, Z, L, W, H, \theta)$, is defined by its 3D location (X, Y, Z), its physical length, width and height (L, W, H) and its orientation θ in the scene. We define the length, width and height (L, W, H) of the proposed 3D object proposals O' as the mean values of annotated 3D objects in the training set. We drop the 3D location coordinate Y since all the 3D object proposals are assumed to be supported by the ground plane, hence $Y = 0$ for all the proposals. Then, once we have generated all the 3D objects that can physically be in the scene, using the camera parameters we project each of the 3D object proposals O' to the image space, assuming a perspective camera model, producing a set of 2D object proposals o'. Specifically, each 2D proposal o' is obtained by projecting each of the corners of the 3D proposal O', and selecting the 2D points that enclose the rest. Note that due to the box representation, objects with opposite orientations (orientation difference = 180°) will project onto the same 2D bounding boxes. For this reason we only generate proposals for a smaller set of $K/2$ discrete viewpoints.

Class independent object proposals

Here we follow the strategy of generic, class-independent, object proposal generators. A crucial part of this strategy is to define a proper objectness measure to be able to estimate how likely it is for a window defined over an image to contain an object of any class. In this analysis we evaluate the effectiveness of this strategy to recover missed detections. Particularly, we use the Selective Search (Uijlings et al., 2013) and Edge Boxes (Zitnick and Dollár, 2014) methods. See Hosang et al. (2014) for a benchmark of methods for detecting class independent object proposals.

3.3. Class-specific context-based object proposals

In this strategy we generate a set of object proposals o' as a function $o' = f^{\eta}(o)$ of the object detections o predicted by the appearance-based detector. The function f^{η} enforces contextual information in the form of relations between object instances. This way, all the proposals o' sampled from f^{η} follow a distribution of relations previously seen in the training data where η is the number of object instances participating in the relation. This produces a relation-driven search where given a seed object o_i object proposals o' are sampled at locations and with poses that satisfy these relations. In this paper we propose two relation-driven functions: f^2 for the case of objects being associated by pairwise relations, and f^+ for the case when objects are associated by higher-order relations.

From 2D object detections to 3D objects in the scene. In this work, reasoning about relations between objects is performed in the 3D scene. For this reason, we first need to project the object detections used as seeds on the 3D scene using the groundplane. We define the objects $O = \{O_1, O_2, \ldots, O_n\}$ as 3D volumes that lie within this 3D space. Each object $O_i = (X_i, Y_i, Z_i, L_i, W_i, H_i, \theta_i, s_i)$, is defined by its 3D location coordinates (X_i, Y_i, Z_i), its size (L_i, W_i, H_i), its pose θ_i in the 3D scene and its confidence score s_i. We assume that all the objects rest on a common ground plane, so $Y_i = 0$ for all the objects. For brevity, we drop the Y term, then each object is defined as $O_i = (X_i, Z_i, L_i, W_i, H_i, \theta_i, s_i)$. In order to define the set of 3D objects O from the set of 2D objects o, we execute the following procedure: first, given a set of annotated 3D objects, we obtain the mean size (length,width and height) of the objects in the dataset. Second, assuming a calibrated camera, we densely generate a set of 3D object proposals O' over the ground plane, very similar the 3D Sliding Windows method from Section 3.2. Third, each of the 3D object proposals from O' is projected in the image plane producing a set of 2D proposals o'. Then, for each object detection o_i we find its corresponding proposal o'_i by taking the proposal with highest intersection over union score, as proposed in Pascal VOC Challenge (Everingham et al., 2012). Finally, we use the 3D location (X'_i, Z'_i) from the 3D proposal O'_i from which o'_i was derived and the viewpoint angle α_i, predicted by the detector, to estimate the pose angle θ_i of the object O_i in the scene. As a result, we obtain a set of 3D objects defined as $O_i = (X_i, Z_i, L_i, W_i, H_i, \theta_i, s_i)$.

Pairwise Relations (f^2). Pairwise relations between 3D objects are computed as proposed in Oramas et al. (2013). Following the procedure from Oramas et al. (2013) we define *camera-centered (CC)* pairwise relations by centering the frame of reference in the camera. Then, from this frame of reference we measure relative location and orientation values between object instances. Alternatively, we define *object-centered (OC)* relations in which, first, we center the frame of reference on each of the object instances and then measure the relative values between them. As a result, we obtain a set of relations R for each image. Each pairwise relation r_{ij} is defined as $r_{ij} = (r_X, r_Z, r_\theta)$, where ($r_X$, r_Z) represent the relative location of the object and r_θ represents the relative pose between the object instances. We compute pairwise relations between each pair of objects within each image of the training set. Then, using kernel density estimation (KDE) we model the distribution $p(r_{ij})$. This is a simple method that manages to find some common arrangements in which pairs of objects co-occur. See Fig. 2 for some examples. During the proposal generation stage, we sample a set of relations r' from this distribution. Then, for each seed object O we generate object proposals O' following the sampled relations r'. Finally, the 3D object proposals O' are projected into the image plane producing the 2D object proposals o'.

Higher-order relations discovery (f^+). Given a set of training images containing objects occurring in a scene, our goal is to discover the underlying higher-order relations that influence the location and orientation in which each object instance occurs w.r.t. each other. A similar problem, of discovering abstract topics $t = \{t_1, t_2, \ldots, t_T\}$ that influence the occurrence of words w within a document d, is addressed by Topic Models (Blei et al., 2003; Griffiths and Steyvers, 2004). Motivated by this similarity we formulate our higher-order relation discovery problem as a topic discovery problem. According to the topic model formulation, a document d_i can cover multiple topics t_k and the words w that appear in the document reflect the set of topics t_k that it covers. From the perspective of statistical natural language processing, a topic t_k can be viewed as a distribution over words w; likewise, a document d can be considered as a probabilistic mixture over the topics t.

In order to meet this formulation in our particular setting, given a set of training images, we first compute pairwise relations r_{ij} between all the objects O_i within each image as before. Then, for each object O_i we define a document d_i where the words w are defined by the pairwise relations r_{ij} that have the object O_i as the source object. Additionally, we experiment with an alternative way to compute the pairwise relations between objects. Specifically, we run tests with a variant of the relative pose attribute of the relation where instead of considering the pose of the target object we consider the orientation of its elongation only (similar to Oramas and Tuytelaars (2014)). This orientation is less affected by errors during prediction, since traditional pose estimators tend to make mistakes by confusing opposite orientations, e.g. front-back, left-right, etc.

In order to make the set of extracted pairwise relations R applicable within the topic model formulation we quantize them into words (although word-free topic models have been proposed as well (Rematas et al., 2012)). To this end, we discretize the space defined by the relations R by ($W/2$, $W/2$, θ_d) where W is the average width of the annotated 3D objects in the training set, and θ_d is a predefined number of discrete poses of the object, 8 in our experiments. At this point, we are ready to perform topic modelling in our data. Here we use Latent Dirichlet Allocation (Blei et al., 2003) for topic modelling. For inference, we follow a Gibbs sampling method as in Griffiths and Steyvers (2004). Our main goal is to identify the set of topics t that define higher-order arrangements between objects O in the scene. In our experiments we extract 16 topics from our documents d. Fig 3 shows a top view of a subset of the discovered topics when considering object-centered pairwise relations as words. Notice how some of the topics resemble common traffic patterns of cars in urban scenes. These topics represent the underlying higher-order relations that we claim influence the way in which objects tend to co-occur.

During the object proposal generation stage, we assume that each 3D object O_i, estimated from the seed object detection o_i, is related with the object proposals O' under higher-order relations. For simplicity, we assume that all the higher-order relations (topics) are equally likely to occur. Object proposals O' are then generated by sampling the word distributions $p(w|t)$ given each of the topics t. Finally the sampled 3D object proposals O' are projected to the image plane, yielding o'. The assumptions made at this stage

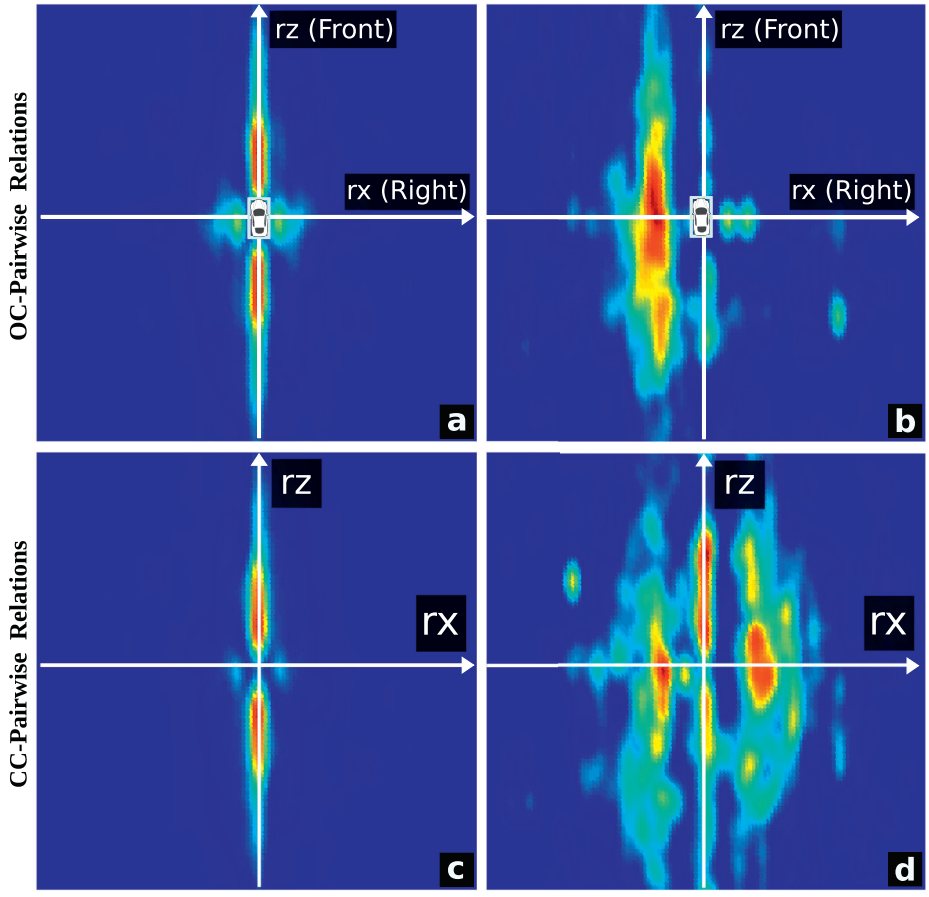

Fig. 2. Distribution of pairwise relations for cars with the same pose (a,c) and opposite pose (b,d) respectively. Top row corresponds to object-centered (OC) relations while the bottom row corresponds to camera-centered (CC) relations.

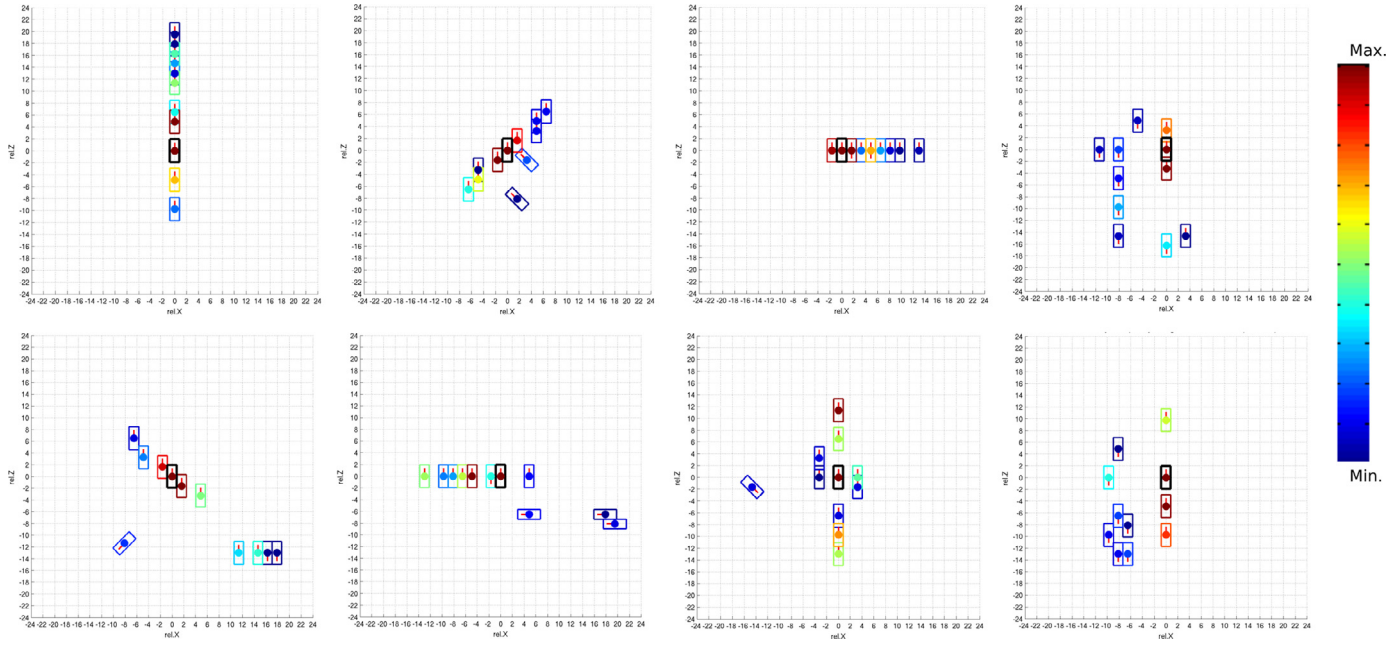

Fig. 3. Some of the discovered relational topics from an object-centered perspective. For each topic, the reference object is in the center and colored in black. The related objects are presented with their occurrence likelihood color-coded in jet scale. Notice how the discovered topics resemble traffic scenarios from urban scenes. For visualization purposes, each object is being plotted with average size of the annotations in the training set of images. We only show the top 10 most likely words per topic. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

have three desirable effects. First, object proposals are sampled in such a way that they follow the higher-order relations between objects. Second, the exploration process gives priority to the most likely proposals from each of the discovered higher-order relations, see Fig. 3. Third, we are able to reason about higher-order relations even for the scenario when just one object detection o_i was collected by the detector.

4. Evaluation

Experiment details: We perform experiments on the KITTI object detection benchmark (Geiger et al., 2012). This dataset constitutes a perfect testbed for our analysis since it covers a wide variety of difficult scenarios ranging from object instances with high occlusions to object instances with very small size. Furthermore, it provides precise annotations for objects in the 2D image and in the 3D space, including their respective viewpoints and poses. In our experiments, and in contrast to standard procedure on the KITT object detection benchmark (Geiger et al., 2012), we consider all the object instances occurring on the images independently of their size, level of occlusion and/or truncation. Since this is a benchmark dataset, annotations are not available for the test set. For this reason, we focus our experiments on the training set. Using the time stamps of the dataset, we split the data into two non-overlapping subsets of equal size. The first subset is used for training, the second subset is used to evaluate the performance of our method. We focus on cars as the class of interest given its high occurrence within this dataset which makes it appropriate for reasoning about relations between objects. We focus our evaluation on images with two or more objects, where it is possible to define such relations. This leaves us with two subsets consisting of 2633 images each that are used for training and testing, respectively. Matching between annotated objects and object proposals is evaluated based on the intersection over union (IoU) criterion from Pascal VOC (Everingham et al., 2012). We report as evaluation metric the recall as a function of the number of object proposals generated per image, as is often used for evaluating object proposal methods. In this analysis we use mainly the LSVM-MDPM-sv detector from Geiger et al. (2011) to collect the initial set of object detections. LSVM-MDPM-sv is an extension of the Deformable Parts-based model (DPM) detector (Felzenszwalb et al., 2010) where a component is trained for each of the discrete object viewpoints to be predicted. In this case, LSVM-MDPM-sv is trained to predict eight viewpoints.

As baselines we use the *Relaxed Score Detector*, the *Relaxed NMS Detector*, the *3D Sliding Window* proposals, the proposals generated by *Selective Search* (Uijlings et al., 2013) and *Edge Boxes* (He and Sun, 2014). For the case of class-specific context-based proposals, we evaluate one method based on pairwise relations, *Pairwise*, and two methods based on higher-order relations, *HOR* and *HOR-Elongation*, where the latter is the variant based on object elongation orientation instead of object pose. For all the context-based strategies, for the special case when no seed objects are available, i.e. images where the object detector was unable to find detections above the threshold (12% of the images), we fallback to the *3D Sliding Window* strategy and consider the proposals proposed by this strategy for that image. We evaluate the changes in performance when considering camera-centered (*CC*) relations vs. object-centered (*OC*) relations.

Exp.1: Relations-based object proposals

In this first experiment we focus on evaluating the strategies based on relations between objects. We consider as seed objects for our strategies the object detections collected with the detector (Geiger et al., 2011). Fig. 4a presents performance on the range of [0,1000] generated object proposals.

Discussion: Strategies based on *CC* higher-order relations seem to dominate the results. They achieve around 10% higher recall than all other methods over a wide range of the curve. This can be attributed to the fact that higher-order relations consider object arrangements with more than two participating objects. This allows them to spot a larger number of areas that are likely to contain objects. In addition, higher-order relations cover a wider neighborhood, whereas the pairwise relations have a more "local" coverage (i.e. they explore mostly a small neighborhood around the seed detections). As a result, strategies based on higher-order relations are able to explore a large part of the image. This is more visible in the range [0,500] of the sampled proposals, where recall from methods based on higher-order relations increases faster than for pairwise relations. This can be further verified in Fig. 5. A deeper inspection of the qualitative results (Fig. 5) produced by our methods reveals a particular trend on how it addresses object instances of different sizes. Our method first focuses on objects in the 3D vicinity of the seed objects, i.e. with similar projected 2D size. Eventually, objects with different 2D sizes to the seeds are explored.

Further we note that strategies based on *CC* relations have superior performance compared to their *OC* counterparts. This can be partly attributed to the fact that proposals sampled following *OC* relations are affected by errors during the prediction of the pose of the seed object. Moreover, the camera setup in the KITTI dataset is fixed, introducing low variability in the *CC* relations. In a scenario with higher variability on camera viewpoints we expect *OC* relations to have superior performance over *CC* relations. In addition, for the case of *CC* relations, the higher-order relations where the elongation orientation is considered are slightly better, albeit only marginally so. This can be attributed to the fact that the orientation of the elongation of an object is less affected by errors in the pose estimation. Moreover, by defining *CC* relations we also avoid the noise introduced in the pose of the seed objects.

Despite the difference in performance between the proposed strategies, it is remarkable that we are able, on average, to double the initial recall obtained by the object detector by following relatively simple strategies. This suggests that object proposal generation should not be employed solely as a pre-detection step as it is commonly found in the literature (Alexe et al., 2010; Endres and Hoiem, 2014; Uijlings et al., 2013; Zitnick and Dollár, 2014). Furthermore, this suggests that there is some level of interoperability between object detection and object proposal generation methods.

Exp.2: Starting from a single object seed

This experiment is similar to the previous experiment with the difference that for each image we only consider the top scoring object detection as seed object. As stated earlier, appearance-based detectors can be reliable at levels of high precision and low recall. The objective of this experiment is to measure what performance can be achieved if we start from the most reliable seed object only. Similar to the previous experiment, Fig. 4b shows performance on the range of [0,1000] generated object proposals.

Discussion: A quick inspection of Fig. 4b shows similar trends as the ones observed in the previous experiment. However, different from the previous experiment, recall is relatively lower in the range of [0,100] proposals. This is to be expected since we start from a smaller pool of seed objects. However, it is surprising to see how we can achieve nearly similar performance from 400 proposals upwards by just starting from a single seed object. This further supports the idea of interoperability between object detectors and object proposal generators.

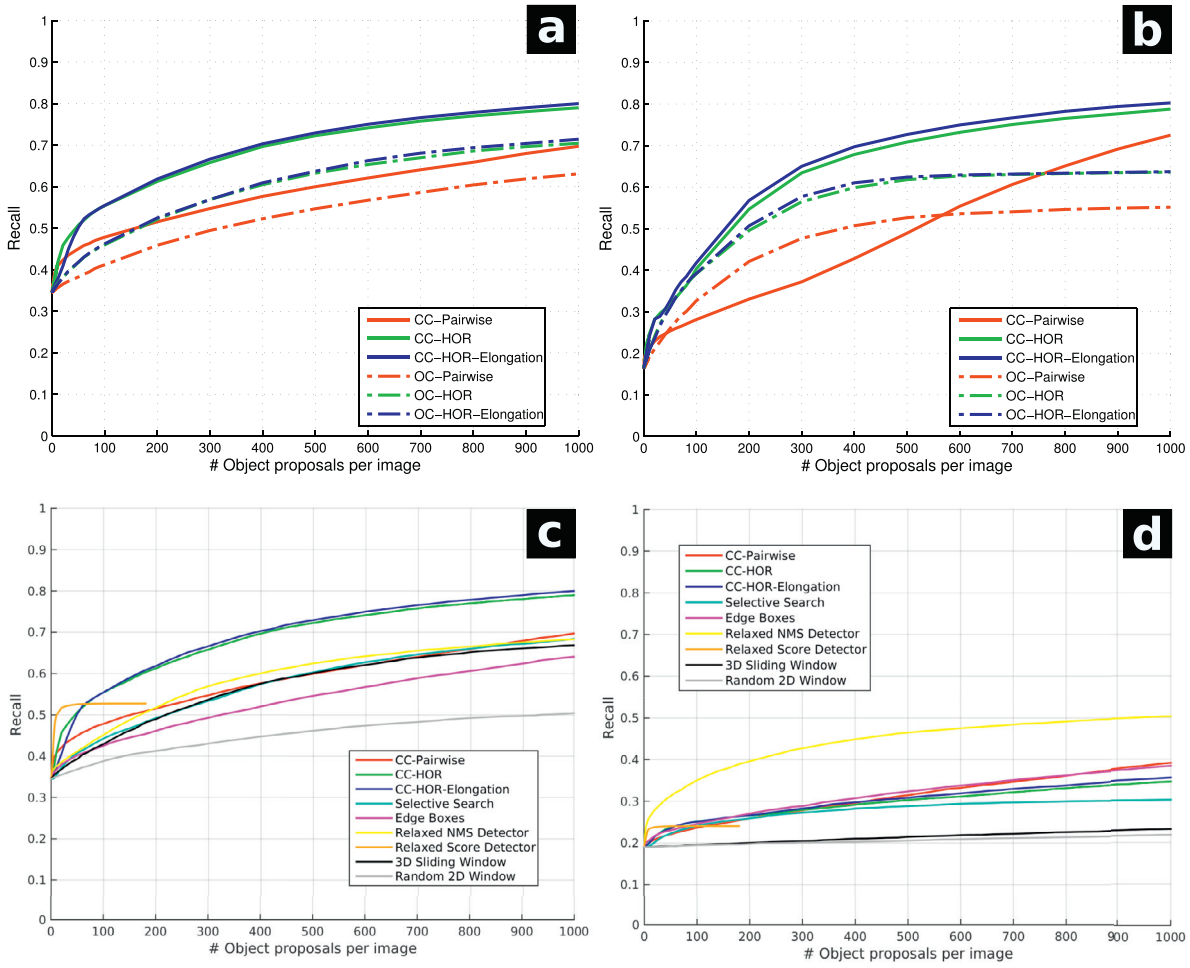

Fig. 4. Recall vs. number of generated proposals for our Relations-based methods a) when all the object detections reported by the detector are used as seed objects, and b) when only the top scoring object detection reported by the detector is used as seed object. Comparison with non-contextual strategies using: c) a traditional matching criterion (IoU $>$ 0.5), and d) using a stricter matching criterion (IoU $>$ 0.75).

Exp.3: Comparison with non-contextual strategies

Next, we compare the performance of the relations-based strategies w.r.t. the non-contextual methods of Section 3.2. We consider as relation-based strategies the *CC* variants only since, in the previous experiments, they achieved higher performance than their *OC* counterparts. As non contextual strategies we consider the *Relaxed Score Detector*, the *Relaxed NMS Detector*, the *3D Sliding Window, Selective Search* (Uijlings et al., 2013), and *Edge Boxes* (Zitnick and Dollár, 2014). We report results considering all the detections as seed objects in Fig. 4c.

Discussion: We notice that the contextual strategies based on higher-order relations have a superior performance than all the other strategies. Interestingly, a clear difference can be noted between the performance of contextual and non-contextual strategies. Except for the *Relaxed Score Detector*, in the range of [0,200], all the contextual strategies achieve superior performance than the non-contextual counterparts. This suggests that indeed contextual information is useful for an early exploration of regions of the image that are likely to host instances of the objects of interest. By observing the performance of the 'Relaxed' versions of our local detector, we can verify the effect that the score threshold and NMS steps have on the obtained recall. As can be noted in Fig. 4c, when reducing the detection score threshold (*Relaxed Score Detector*), the set of hypotheses predicted for each image is much lower ($<$ 200 per image) than when the NMS step is reduced (*Relaxed NMS Detector*). Due to its stricter NMS step, the *Relaxed Score Detector* produces less overlapping hypotheses, hence performing a faster exploration of the image space. This is evident since it reaches relatively high recall ($\sim$ 0.5) at the cost of less than 50 proposals per image. On the downside, due to its limited number of predicted hypotheses, this recall is not able to increase significantly. On the contrary, when the NMS step is removed, the *Relaxed NMS Detector* reaches the 0.5 recall of the *Relaxed Score Detector* at $\sim$ 200 proposals per image and is able to reach up to a recall of 0.7 later on the curve.

Exp.4: Proposal localization/fitting quality

In this experiment we measure the quality of the object proposal to localize and fit the region of the recovered object instance. For this purpose, in this experiment we employ a stricter matching criterion (Everingham et al., 2012) of at least 0.75 IoU between the bounding boxes of the object annotations and the object proposals, respectively. This is inspired by Hosang et al. (2014), where it is claimed that a 0.5 IoU is insufficient for evaluating object proposals. We evaluate the performance of the same, contextual and non-contextual, strategies from Exp.3. In Fig. 4d we report results considering all the detections as seed objects with stricter IoU measure.

Fig. 5. Object proposals generated in chronological order using the context-based strategy based on camera-centered higher-order relations. Matched object annotations are marked in blue, missed detections are marked in red and matching object proposals are color-coded in green. First row, seed objects collected with the object detector (Geiger et al., 2011); second row, results after sampling 100 object proposals; and third row, results after sampling 1000 object proposals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Discussion: Recall values obtained in this experiment are significantly reduced now that matching an object is a more complicated task. The performance of the *Relaxed NMS Detector* is surprisingly high. This can be attributed to the fact that with the non-maximum suppression step removed the *Relaxed NMS Detector* is able to exhaustively explore the areas where appearance has triggered a detection. This is further confirmed when comparing its performance with the one of the *Relaxed Score Detector*. Due to its stricter NMS step, the *Relaxed Score Detector* is not able to explore possible bounding box variants occurring on a candidate region, thus resulting in poorer hypothesis bounding box matching. In addition, we notice that pairwise relations are now outperforming the higher-order alternatives in the range of [500,1000] proposals/image. This may be caused by the discrete nature of the words in the topic models which are used to discover higher-order relations. As a result, the proposals generated from higher-order relations are spatially sparser than the ones produced by pairwise relations. The strategies based on pairwise relations tend to first concentrate on regions of high density before exploring other areas. This is why we notice improvements in the range [500,1000] and not earlier. These observations hint at a possible weakness of our relation-based strategies to generate object proposals. On one hand, relation-based proposals have some level of sparsity embedded, in our case, either by vector quatization of the relational space or by assuming mean physical sizes for the objects in the scene, when reasoning in 3D. This can be a weakness compared to the

Table 1
Detection performance. Mean Average Precision (mAP).

Baseline	Raw	SPP-CNN	
		SVM class.	bbox regress.
3D Sliding Window	31.30	21.11	33.53
Random 2D Windows	30.97	18.00	24.63
Selective Search	31.40	33.50	43.50
Edge Boxes	31.29	35.12	39.63
CC-Pairwise	36.37	21.44	31.19
CC-HOR	**37.68**	34.20	**48.19**
CC-HOR-Elongation	36.25	34.30	**48.40**
Relaxed NMS Detector	34.40	**42.77**	**48.27**

exhaustive *Relaxed NMS Detector* strategy, when the objective is to have fine localization. On the other hand, relation-based strategies seem to be better suited for "spotting" the regions where the objects of interest might be. This is supported by their superior recall in Exp.1. This further motivates our idea of a joint work of object detectors and object proposal generators.

Exp.5: Measuring detection performance

While the results obtained in the previous experiments show a significant improvement in recall, it is arguable whether the cost of decreased precision is acceptable. We argue that, in systems with various sources of information (multimodal sensors, multi-cameras or image sequences) it is desirable to detect the majority of the objects since the pool of detections can be further reduced by imposing consistency along the different sources. In order to get a notion of the potential of the proposed methods for the object detection task, we will now follow the traditional object detection evaluation protocol. We report Mean Average Precision (mAP) as performance metric and use the standard matching criterion from Pascal VOC (IoU $>$ 0.5). First, we report the performance for the *raw* set of object proposals (1000 objects/image) from the previous experiments. Second, aiming at reducing the number of false positives per image and at having a comparison w.r.t. state-of-the-art detection methods, we re-score the *raw* set of objects using appearance features. For this purpose, we follow the R-CNN strategy (Girshick et al., 2014). Given a set of object proposals, we compute CNN features for each proposal and then classify each region using a linear SVM. As an additional step, linear regression is performed in order to fix bounding box localization errors. In this experiment we consider the set of objects from the previous experiments as the proposals to be classified. Finally, we follow the SPP-CNN alternative from He and Sun (2014) which has comparable detection performance to R-CNN at a fraction of processing speed. In addition, we split the performance of our CNN-based baselines showing the performance after SVM classification and after performing bounding box regression, respectively. See Table 1 for some quantitative results. In Fig. 6 we present the precision-recall curves. Following the experimental protocol presented in this section, we present performance curves for SVM classification (Fig. 6a) and for bounding box regression (Fig. 6b). See Fig. 7 for some qualitative examples from this experiment.

Discussion: We notice that when the set of raw objects is considered, our relations-based methods lead the performance table. This group of methods is followed by the *Relaxed NMS Detector*, the class-independent methods and the random methods, respectively. This shows that at this "raw" level, the proposed relations-based methods are better suited to cope with the variations in object appearance caused by occlusions and changes on scale and viewpoint. For the case when appearance-based re-scoring is performed, it is important to notice that the combination of SPP-CNN with *Selective Search* proposals corresponds to the state-of-

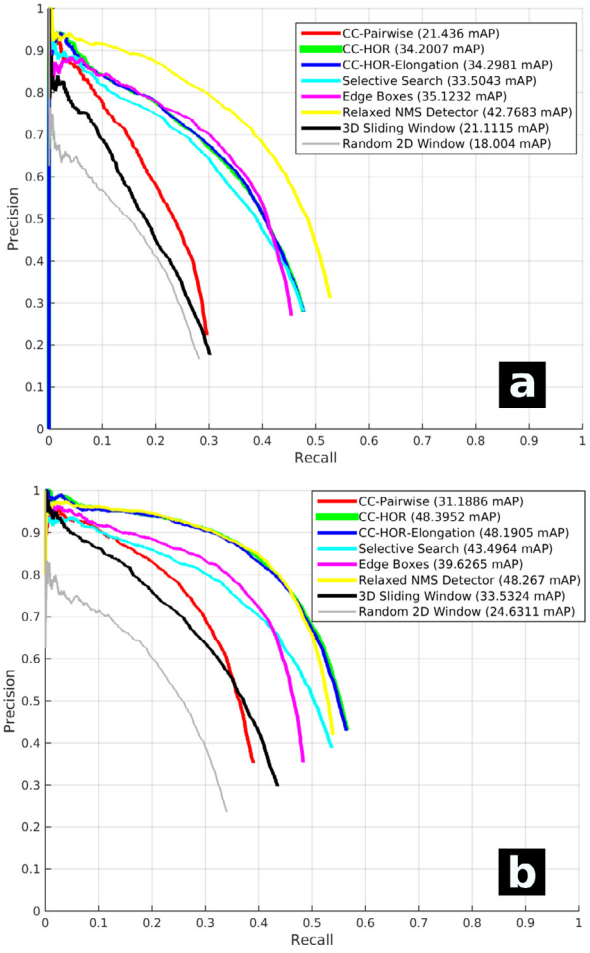

Fig. 6. Object detection Mean Average Precision (mAP) performance of CNN-based methods. Precision-Recall curves showing the performance obtained a) after SVM classification, and b) after bounding box regression.

the-art method proposed in He and Sun (2014). Furthermore, as mentioned earlier, this is a speeded-up version of R-CNN (Girshick et al., 2014). We can notice that the proposed relations-based methods produce an improvement of $\sim$ 5 percentage points (pp) over R-CNN (Selective Search + SPP-CNN). Furthermore, the comparable good results based on the *Relaxed NMS Detector* proposals suggests that using a weaker detector as proposal generator can boost the results obtained with SPP-CNN features. While not at the core of our paper, this seems an interesting observation. In addition, when comparing the difference in performance within the SPP-CNN setup, it is clear that our relations-based methods benefit more from bounding box refinement ($\sim$ 13 pp improvement) than the appearance-based *Relaxed NMS Detector* ($\sim$ 7 pp). This further confirms our observation made in the previous experiments that context-based proposals are better at spotting regions of the image likely to contain the objects while appearance-based approaches are better suited for finer localization. An additional difference between the performance of the *Relaxed NMS Detector* and the relations-based methods lies in their processing times. In their current state, the evaluated baselines, e.g. *Relaxed NMS Detector* and the relations-based methods, have a processing bottleneck in the way in which the seed hypotheses are obtained. Since for the *Relaxed NMS Detector* the set of hypotheses is high ($\sim$ 1000 hypotheses), its processing time is much higher than the relations-based methods which usually start from a set of $\sim$ 7 seed object hypotheses. This difference in processing times will be further discussed in Section 5.

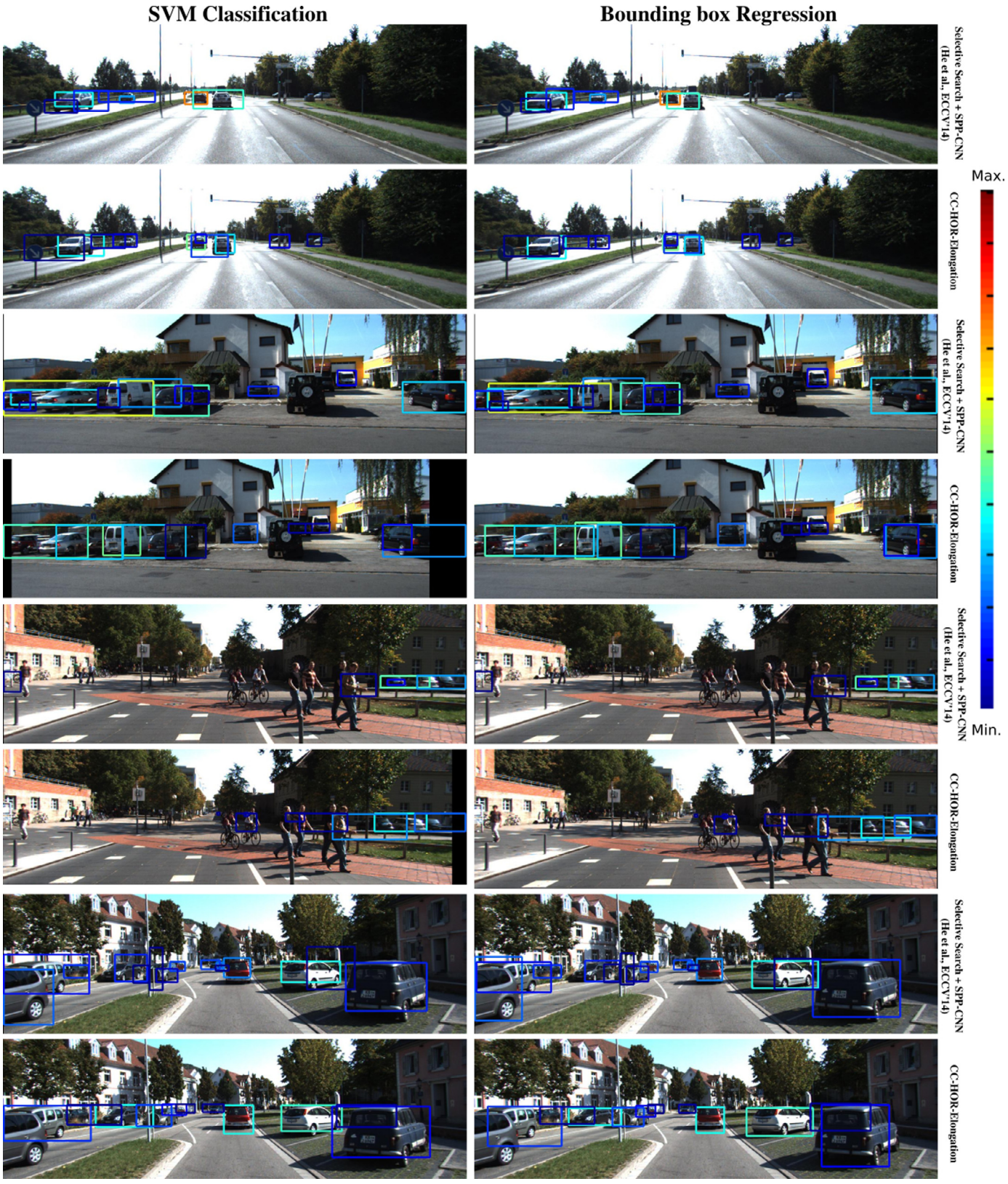

Fig. 7. Qualitative examples from the CNN-based methods for object detection. For each image the detection score of each hypothesis is color-coded in jet scale. We show the examples for the methods from (He and Sun, 2014) (SPP-CNN + Selective Search (Uijlings et al., 2013)) and the proposed camera-centered higher-order relations method (CC-HOR-Elongation). We show the hypotheses predicted a) after SVM classification, and b) after bounding box regression. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Exp.6: Comparison w.r.t. the state-of-the-art

In order to further evaluate the strength of our context-based methods at recovering missed object instances, in this experiment we evaluate its performance when starting from a state-of-the-art method for object detection, i.e. the Faster R-CNN detector (Ren et al., 2015). In this experiment we define three methods, the *Vanilla Faster R-CNN detector* which is the original version of the detector as proposed in Ren et al. (2015), with default parameters for detection score threshold and non-maximum suppression. As second set of method s we have the relaxed versions (*Relaxed Score* and *Relaxed NMS*) of the detector. Finally, we have our context-based methods where each of the hypotheses produced by the *Vanilla R-CNN detector* are enriched with viewpoint predictions using a multiclass SVM classifier trained from CNN features (Jia et al., 2014) computed from annotated instances in the dataset. Something important to note, is that this classifier is not perfect: it achieves a training cross validation accuracy of 0.4. However, its performance is above chance levels so it can give us an idea of the viewpoint (or at least the elongation angle) of an object. Similar to the experiments reported earlier, we used the hypotheses collected by the Faster RCNN detector, with default settings, as seed objects. Based on these hypotheses we sample context-based object proposals. Similar to the previous experiments, for the case when no seed objects are available, i.e. images where the object detector was unable to find detections above the threshold (1% of the images), we fallback to the *3D Sliding Window* strategy and consider the proposals proposed by this strategy for that image. We report performance in terms of Recall as a function of the number of sampled object proposals (see Fig. 8).

Discussion: Clearly, the Faster RCNN detector achieves better performance (achieving an initial Recall $\sim$ 0.55) than the DPM-based detector used earlier in our experiments (which achieved an initial Recall $\sim$ 0.35). As a result, the proposed context-based methods are fed with better seed objects resulting in a boost in performance (now being able to reach a recall of $\sim$ 0.9). Note that, as was stressed earlier, this is achieved by using noisy object viewpoint estimates. We believe that a state-of-the-art method for viewpoint estimation, e.g. Choy et al. (2015), can help to boost the performance of the proposed context-based methods further. In this experiment we also note the same trend on the performance of the different methods when using a stricter intersection over union (IoU) matching criterion (Fig. 8b).

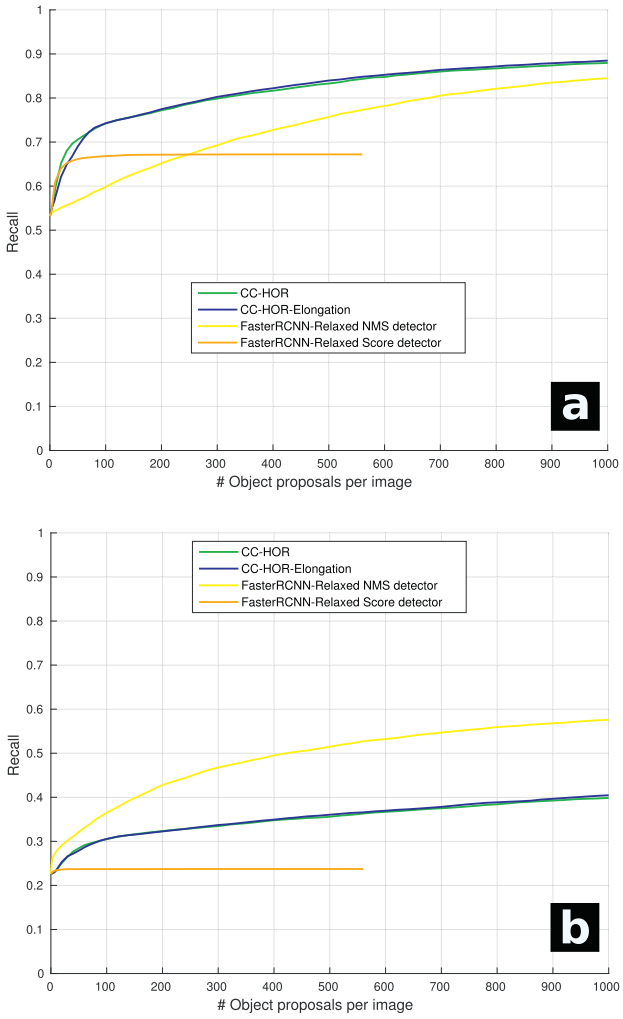

Fig. 8. Recall vs. number of generated proposals for our Relations-based methods when seed objects are collected with the Faster RCNN detector (Ren et al., 2015) at default settings. For reference, we also report the performance of the *Relaxed Score/NMS* detectors. We report performance when using: a) a traditional matching criterion (IoU $>$ 0.5), and b) using a stricter matching criterion (IoU $>$ 0.75).

5. Processing times

Regarding processing times, each of the methods in Table 1 considers the same number of proposals/image. This leads to similar processing times during the appearance-based re-scoring (classification/regression). Hence, the difference in the processing time between the evaluated methods is determined by their respective methods to generate object proposals. As mentioned earlier, the proposal generation process of the proposed method consists of two steps: a) class-specific object seed detection, and b) context-based proposal generation.

In its current state, the bottleneck of the proposed method lies in the seed detection step which is handled by an off-the-shelf detector (Geiger et al., 2011). Class-specific object seed detection takes on average 20 s/image when using the detector at default settings. For the second step, given that the context models have been computed offline (Section 3.2), the execution of the proposed method can be summarized into three main processes, i.e. 2D-3D projection, topic assignment and object proposal sampling; which all scale linearly w.r.t. the number of desired object proposals. These processes take approximately 0.5, 0.1 and 1 s/image, respectively, giving a total of 1.6 s/image for the sampling of context-based proposals.

For the case of the *Relaxed NMS Detector*, the object detector needs to be evaluated for a large set of windows. Note that compared to the proposed method, the number of hypotheses collected by the off-the-shelf detector is very high in the *Relaxed NMS Detector*. This increases the processing within the off-the-shelf detector (Geiger et al., 2011) to 30 s/image and further increases the bottleneck mentioned earlier.

Note that the computation times presented above are obtained by performing only CPU-based computations. Moreover, for both the proposed method and the *Relaxed NMS Detector*, this problem can be alleviated by using faster detectors, e.g. Benenson et al. (2012); He and Sun (2014); Ren et al. (2015). This is supported by our experiments based on the Faster RCNN detector (Ren et al., 2015) (Section 4, Exp.6) where on the one hand the detection of seed objects takes $\sim$ 0.17 s/image while on the other hand the *relaxed NMS* detector takes $\sim$ 1.12 s/image.

6. Limitations and future work

Even though we have shown that the proposed method is effective at recovering object instances missed after an initial detection

step, there are several aspects in which the proposed method can be improved. In this section we look at these weak points and suggest directions for addressing them in future work. Currently, our evaluation is focused on grounded objects of a single-class, i.e. the car class on a groundplane. Even though, our relations-based models can be extended to cover other classes not necessarily on the groundplane, e.g. by adding the relative Y location r_Y and related object class r_C as relation attributes, making this extension comes with the cost of requiring additional training data. As was presented in Section 3.3, the proposed methods to generate context-based proposals learn relations between objects from training data. Thus, as the definition of pairwise relations gets more complex, more representative training data would be required in order to cover all the new scenarios that might be possible with the new extended pairwise relations model. In this regard, further experiments should be performed to verify the performance of these extended models. In addition, by being class-specific, our method may not scale properly if a large number of object classes need to be detected. In this regard, we suggest the usage of our method for structured scenarios with a reduced number of classes, e.g. autonomous driving and indoor object detection, or as a detector for specific scene-types. Considering a specific type of setting or scene, will reduce the number of object classes that need to be analyzed during test time making the scalability aspect less critical.

Regarding the object seed detection step, in its current state, our method requires a detector that provides a viewpoint as part of its output. This requirement can be alleviated by modifying the relations-based models to focus on spatial relations (ignoring the relative pose information). This change comes at the cost of less interpretable context models. A more promissing solution follows the recent line of work from Choy et al. (2015); Rematas et al. (2015) which focuses on registering 3D (CAD) models to objects depicted in 2D images. As is presented in Choy et al. (2015), this registration can be successfully exploited to enrich detected object hypotheses (bounding boxes) with information related to viewpoint. Given the increasing amount of 3D models appearing everyday, methods like Choy et al. (2015) clearly address the requirement of having object detections with predicted viewpoint. Moreover, as was presented on Exp.6 (Section 4), even when using a relatively simple, and noisy, viewpoint estimator (CNN features+SVM) decent performance can be achieved by the proposed method.

As presented in Fig. 4d, when focusing on fine object localization, using an exhaustive dense *Relaxed NMS Detector* outperforms the proposed method. In order to improve the performance of the proposed method on the fine localization task, inspired by Long et al. (2014), we propose to follow a top-down approach in which given a set of object seeds we generate a set of initial relations-based proposals from which proposals with controlled variations are sampled. Size and location of these additional proposals are ruled by statistical data related to the class of interest and the spatial location of other objects in the scene.

Regarding the assumptions made on the proposed method, having a calibrated camera might sound as a strong assumption. However, note that existing works (Bao et al., 2011; Hoiem et al., 2006; Wang et al., 2005; Wilczkowiak et al., 2001) have proposed several methods to perform this calibration.

In this work we have focused our evaluation on the KITTI dataset (Geiger et al., 2012). As stated in Section 4, this dataset constitutes a perfect testbed for our analysis since it covers a wide variety of difficult scenarios, e.g. object instances with high occlusions, object instances with very small size, etc. Furthermore, it provides precise annotations from objects in the 2D image and in the 3D space, including their respective viewpoints and poses. Finally, most of the images of the KITTI dataset contain more than one instance of the class of interest, i.e. car, which is necessary for learning the relations between objects. As future work, further evaluation of the proposed method should be performed as new datasets showing similar properties to the KITTI dataset appear.

Finally, comparing the performance of both the *Relaxed Score* and *Relaxed NMS* detectors, suggests that a proper balance between their thresholds, i.e. NMS and detection score threshold, can be obtained in order to improve detection performance. This somehow goes against the common practice of focusing on the detection score threshold alone and leaving NMS as a fixed post-processing step. Moreover, a good balance between the two relaxed methods may produce a better set of seed objects for our context-based method for generating proposals.

7. Conclusions

In this paper we have shown that sampling class-specific context-based object proposals is an effective way to recover missed detections. The potential of our method to improve detection is shown by our straightforward CNN extension which achieves improved performance over state-of-the-art CNN-based methods. Our experimental results suggest that object proposal generation should not be employed solely as a pre-detection step as it is commonly found in the literature. Furthermore, we show relations-based strategies are better suited for spotting regions that contain objects of interest rather than achieving fine localization. In addition, our novel method to discover higher-order relations is able to recover semantic patterns such as traffic patterns found in urban scenes. Future work will focus on investigating the complementarity of the proposed strategies as well as proper ways to integrate them.

Acknowledgments

This work is supported by the FWO project “Representations and algorithms for the captation, visualization and manipulation of moving 3D objects, subjects and scenes”, and an NVIDIA Academic Hardware Grant.

References

Alexe, B., Deselaers, T., Ferrari, V., 2010. What is an object? CVPR.
Alexe, B., Deselaers, T., Ferrari, V., 2012a. Measuring the objectness of image windows. TPAMI 34 (11), 2189–2202.
Alexe, B., Heess, N., Teh, Y.W., Ferrari, V., 2012b. Searching for objects driven by context. NIPS.
Atanasoaei, C., McCool, C., Marcel, S., 2010. A principled approach to remove false alarms by modelling the context of a face detector. BMVC.
Bao, S.Y., Sun, M., Savarese, S., 2011. Toward coherent object detection and scene layout understanding. Image Vision Comput. 29 (9), 569–579.
Bao, S.Y., Xiang, Y., Savarese, S., 2012. Object co-detection. ECCV.
Benenson, R., Mathias, M., Timofte, R., Van Gool, L., 2012. Pedestrian detection at 100 frames per second. CVPR.
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. JMLR 3, 993–1022.
Cao, X., Wei, X., Han, Y., Chen, X., 2015. An object-level high-order contextual descriptor based on semantic, spatial, and scale cues. Trans. Cybern. 45 (7), 1327–1339.
Choi, M., Lim, J.J., Torralba, A., Willsky, A.S., 2010. Exploiting hierarchical context on a large database of object categories. CVPR.
Choy, C., Stark, M., Corbett-Davies, S., Savarese, S., 2015. Enriching object detection with 2d-3d registration and continuous viewpoint estimation. CVPR.
Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. CVPR.
Desai, C., Ramanan, D., Fowlkes, C.C., 2011. Discriminative models for multi-class object layout. IJCV 95 (1), 1–12.
Endres, I., Hoiem, D., 2014. Category-independent object proposals with diverse ranking. TPAMI 36 (2), 222–234.
Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A. The PASCAL visual object classes challenge 2012 (VOC2012) results.
Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part-based models. TPAMI 32 (9), 1627–1645.
Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. CVPR.
Geiger, A., Wojek, C., Urtasun, R., 2011. Joint 3d estimation of objects and scene layout. NIPS.
Ghodrati, A., Diba, A., Pedersoli, M., Tuytelaars, T., Van Gool, L., 2015. Deepproposal: hunting objects by cascading deep convolutional layers. ICCV.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR.
Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Procs. Natl. Acad. Sci. 101, 5228–5235.
He, Z.X.R.S., Sun, KaimingJ., 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. ECCV.
Hoiem, D., Efros, A.A., Hebert, M., 2006. Putting objects in perspective. CVPR.
Hoiem, D., Efros, A.A., Hebert, M., 2011. Recovering occlusion boundaries from an image. IJCV 91 (3), 328–346.
Hosang, J., Benenson, R., Schiele, B., 2014. How good are detection proposals, really? BMVC.
Hou, Y., He, L., Zhao, X., Song, D., 2011. Pure high-order word dependence mining via information geometry. Adv. Information Retrieval Theory.
Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: convolutional architecture for fast feature embedding. MM.
Kuo, W., Hariharan, B., Malik, J., 2015. Deepbox: learning objectness with convolutional networks. ICCV.
Küttel, D., Breitenstein, M.D., Gool, L.J.V., Ferrari, V., 2010. What's going on? Discovering spatio-temporal dependencies in dynamic scenes. CVPR.
Li, B., Wu, T., Zhu, S.-C., 2014. Integrating context and occlusion for car detection by hierarchical andor model. ECCV.
Li, H., Chen, L., 2010. Removal of false positive in object detection with contour-based classifiers. ICIP.
Long, C., Wang, X., Gang Hua, M.Y., Lin, Y., 2014. Accurate object detection with location relaxation and regionlets re-localization. ACCV.
Mathias, M., Benenson, R., Timofte, R., Van Gool, L., 2013. Handling occlusions with franken-classifiers. ICCV.
Mottaghi, R., Chen, X., Liu, X., Cho, N.-G., Lee, S.-W., Fidler, S., Urtasun, R., Yuille, A., 2014. The role of context for object detection and semantic segmentation in the wild. CVPR.
Oramas, M.J., De Raedt, L., Tuytelaars, T., 2013. Allocentric pose estimation. ICCV.
Oramas, M.J., De Raedt, L., Tuytelaars, T., 2014. Towards cautious collective inference for object verification. WACV.
Oramas, M.J., Tuytelaars, T., 2014. Scene-driven cues for viewpoint classification of elongated object classes. BMVC.
Pepik, B., Gehler, P.V., Stark, M., Schiele, B., 2012. 3d2pm - 3d deformable part models. ECCV.
Pepik, B., Stark, M., Gehler, P.V., Schiele, B., 2013. Occlusion patterns for object class detection. CVPR.
Perko, R., Leonardis, A., 2010. A framework for visual-context-aware object detection in still images. CVIU 114 (6), 700–711.
Rematas, K., Fritz, M., Tuytelaars, T., 2012. Kernel density topic models: visual topics without visual words. NIPS Workshops.
Rematas, K., Nguyen, C., Ritschel, T., Fritz, M., Tuytelaars, T., 2015. Novel views of objects from a single image. arXiv:1602.00328 [cs.CV].
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. NIPS.
Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M., 2013. Selective search for object recognition. IJCV 104 (2), 154–171.
Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A., 2009. Multiple kernels for object detection. ICCV.
Wang, G., Tsui, H.-T., Hu, Z., Wu, F., 2005. Camera calibration and 3d reconstruction from a single view based on scene constraints. Image Vision Comput. 23 (3), 311–323.
Wilczkowiak, M., Boyer, E., Sturm, P., 2001. Camera calibration and 3d reconstruction from single images using parallelepipeds. ICCV.
Yao, B., Li, F., 2010. Modeling mutual context of object and human pose in human-object interaction activities. CVPR.
Zhang, H., Geiger, A., Urtasun, R., 2013. Understanding high-level semantics by modeling traffic patterns. ICCV.
Zia, M.Z., Stark, M., Schindler, K., 2014. Are cars just 3d boxes? - jointly estimating the 3d shape of multiple objects. CVPR.
Zitnick, C., Dollár, P., 2014. Edge boxes: locating object proposals from edges. ECCV.