Towards Cautious Collective Inference for Object Verification

José Oramas M.
KU Leuven, ESAT-PSI, iMinds

Luc De Raedt
KU Leuven, CS-DTAI

Tinne Tuytelaars
KU Leuven, ESAT-PSI, iMinds

Abstract

It is by now generally accepted that reasoning about the relationships between objects (and object hypotheses) can improve the accuracy of object detection methods. Relations between objects allow to reject inconsistent hypotheses and reduce the uncertainty of the initial hypotheses. However, most methods to date reason about object relations in a relatively crude way. In this paper we propose an alternative using cautious *inference. Building on ideas from Collective Classification, we favor the most confident hypotheses as sources of contextual information and give higher relevance to the object relations observed during training. Additionally, we propose to* cluster *the pairwise relations into relationships. Our experiments on part of the KITTI data benchmark and the MIT StreetScenes dataset show that both steps improve the performance of relational classifiers.*

1. Introduction

Recently, contextual information has been used in several computer vision tasks including segmentation [9, 11, 12] and object detection [5, 13, 24]. For the object detection problem, relations between object instances have been used to remove or reduce the uncertainty in hypotheses predicted by appearance-based detectors. A common pipeline in these works proceeds as follows: 1) an initial set of object hypotheses is obtained using an object detector; 2) for each hypothesis, a set of neighbor objects is selected as sources of contextual information; and 3) information from these neighboring objects is used to re-evaluate the initial object. The underlying methods differ in the way they define neighboring objects. Some works (e.g. [5, 23]) use all the other objects as neighbors, while others (e.g. [9]) use only the objects in a spatial vicinity. We refer to these two types as "global" and "near" neighborhoods, and empirically evaluate which setting yields best results.

For inference, the neighboring object hypotheses are commonly considered without taking into account the certainty of their prediction. As as result, *all* the neighbors participate for the classification of each object [23]. Following the literature [17, 18] on Collective Classification [25], instead, we propose an iterative scheme where we first classify the objects with most certain relational information, and then use these to bootstrap the predictions of the other objects. This is useful in collective classification tasks, like object detection, where multiple possibly related objects all need to be classified. Following the terminology of [17], we refer to these two inference variants as "aggressive" or "cautious" inference. Again, we empirically evaluate the added value of cautious vs. aggressive inference.

Furthermore, probabilities or likelihoods are typically computed based on the frequency of occurrences of object relations in the training data. Usually, this is computed relative to *all* the relations involving two objects of the same class. This is an example of classical *homophily*-based relational classification. *Homophily* is the tendency of individuals to associate with others of the *same* class. This homophily-based model is inspired by observations in a vast array of network studies, e.g. [19], in both explicitly defined and latent-assumed networks. In *homophily*-based relational classification, objects are expected to give higher support to hypotheses belonging to the same class [16] independent of the relation between them. Here we also investigate an alternative definition for *homophily*, based on the relation between object instances rather than strictly focused on the classes of the objects. Following this idea, we assume that the observed pairwise relations between objects belong to a set of underlying relationships that determine how the different objects are associated with each other. In this setting, during inference, only a subset of the relations (those covered by the same relationship) are involved in the estimation of probabilities or likelihoods. We refer to these two cases as "class-based homophily" and "relation-based homophily", and empirically evaluate their respective merits. Let us illustrate these ideas by an example. Imagine you are given the task of predicting whether the green box in Fig. 1 (corresponding to an object hypothesis) contains a car or not, based on the context given by the objects in the other bounding boxes (Fig. 1a). Shouldn't the true hypotheses, in blue, have a higher influence on the prediction than the false hypotheses, in red? Furthermore, focusing on the true hypotheses (Figure 1b), wouldn't it be more intuitive to take into account also the color-codes of the objects?

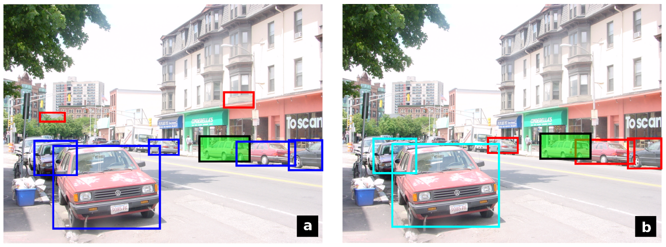

Figure 1. Different ways in which neighboring objects can cast a vote for the query object (green box): a) cautious inference, selecting the more reliable neighbors, b) relation-based homophily, considering the relations between objects. (See text for details.)

(defined by their relations with the object under the green box). These two figures, Fig. 1a and Fig. 1b , depict examples of cautious inference and relation-based homophily respectively.

The main contributions of this paper are: a) we show that cautious inference on object relations brings improvements over traditional inference for object detection, and b) we investigate a notion of relational homophily, which specifies how neighboring objects influence unknown objects based on underlying relationships. Additionally, we evaluate these techniques using three different representations for object relations.

This paper is organized as follows: Section 2 presents existing work on using relations between objects for object detection. Section 3 introduces how we define and learn relations between object hypotheses provided by a detector and emphasizes the principles to be considered when performing cautious inference with such relations. Section 4 presents how the output of the local and relational classifiers are combined. In section 5 we provide implementation details, while section 6 describes the experimental results. Finally, we draw conclusions in section 7.

2. Related Work

In the literature, sources of contextual cues used for the object detection problem have been divided into three categories [8, 11, 22]: *Things*, i.e. entities with defined appearance and shape; *Stuff*, i.e. entities without a defined appearance and shape; and *Scene*, which is derived from holistic features of the whole image. In this work, we focus on the contextual cues provided by *Things*. To this end, [5] represented objects as regions in the image. Then, by learning qualitative spatial relations (e.g. top-left, far-left) between them, hypotheses in unlikely areas were filtered out. Following this trend, Felzenszwalb et al. [7], Perko & Leonardis [23] and Choi et al. [2] defined continuous spatial relations between the centers of object bounding boxes and used the learned relations to filter out the out of context objects. Song et al. [26] suggested a joint detection-classification scheme to identify ambiguous ranked hypotheses and use an adaptive method to exploit context information on these hypotheses. This method favored the prediction of the local detector on top and low ranked hypotheses and considered the context information for the ambiguous cases. In [3] relations between objects considered additional features such as relative scales, bounding box overlap ratio and scores. Furthermore through a set-based formulation this method is able to reason about object spatial configurations that go beyond pairwise interactions. Similar to these works, we learn relations between object instances. In particular, we build on the work of [23] and their framework based on Kernel Density Estimation (KDE) to estimate the probabilities of certain relations. Different from these works, which follow an “aggressive” approach considering all the neighboring objects as sources of contextual support, we explore an alternative “cautious” inference scheme in which the set of neighboring objects is defined based on the level of certainty of their occurrence. Our method is inspired by *Cautious Inference* [18, 21], a method used in collective classification that seeks to identify and exploit the more certain relational information. Such a cautious approach was used in [12] for the problem of labeling object superpixels. In [12], discriminative relations were mined between object regions and discriminative attributes were discovered per relation. Our approach to object detection differs from that in [12] in that in object detection the object regions (the bounding boxes) can overlap, which increases the complexity of the problem. To the best of our knowledge, this is the first attempt to use *cautious* collective inference to improve object detection.

An important issue in collective classification problems is the way in which each of the neighboring objects casts a vote on the unknown object. Experiments [15, 16] with Collective Classification algorithms on text data used “class-based” homophily models, that is, an object is only associated with objects of the same class. These models have the limitation that the association of objects is strictly class-driven, and the weakness that all the neighboring objects influence directly the occurrence of others of the same class in a direct and rather crude fashion. Recent work in object detection that exploits object relations takes into account inter/intra class object relations, however they still perform inference using the relations between objects directly. We follow the suggestion of [20], which aims to uncover underlying groups, which represent the cause of the frequently seen pairwise relations. However, different from [20], we do not require fully visible explicit relations at both training and testing stages. Furthermore, since in our work we define attribute-based relations, there is a relation between any pair of objects, linked to its underlying group. This removes the requirement of group membership for objects that are not explicitly related. Summarizing, our work differs from the class-based homophily models since we assume that objects are not linked by their class, but by the relationships underlying the visible pairwise relation between objects of the same class. Focusing on the relational aspect of the problem, we formulate our object classification problem as a Within-Network classification problem, which consists of making a prediction about an object based on the neighboring objects.

A recent group of works [13, 24] promotes the use of groups of objects with consistent relations among them. Following this idea, [24] exploits explicitly defined pairwise relations to learn the collective appearance of object pairs. Taking this idea further, [13] removes the requirement of explicitly defined relations and discovers composite relations to learn the appearance of the group of objects. Our procedure of recovering underlying *relationships* and their corresponding densities is quite similar to the Hough transform and mode finding approach in [13]. However, different from [13] and [24], which use underlying groups towards learning the collective appearance of the groups, we discover the underlying groups as a means to improve object detection accuracy.

3. Object Relations as Source of Context

Before we discuss how relations between objects can be used as a source of contextual information, we introduce the representations for objects and relations used in this paper. Given an image, we use an object detector to collect a set of object hypotheses $O = \{o_1, o_2, ..., o_n\}$ of the class of interest. Each object hypothesis o_i is represented as a tuple $o_i = (x_i, y_i, f_i, s_i)$ where (x_i, y_i) represents the location of the center of the bounding box of the object, f_i represents additional object-related features (e.g. aspect ratio or scale of the bounding box), and s_i the detection score reported by the detector. Given the set of hypotheses O, we define pairwise relations r_{ij} between each pair of objects o_i and o_j. In section 5 we describe how we compute the relative attributes that define the relations r_{ij} .

3.1. Inference

In this paper we follow the principle proposed in [17] that stresses that instances are not independent, on the contrary, "in some classification tasks they are implicitly or explicitly related". Therefore, we estimate the degree to which an object o_i fits into the scene based on its relations with the other objects in the scene. This is a *Collective Classification* [25] problem in which the occurrence (class) of an object influences that of another. For simplicity we focus on the case of a single object class for now. To take into account the interdependencies between objects based on their relations we re-rank the predicted object hypotheses using the Weighted Vote Relational Neighbor Classifier (wvRN) [16]. wvRN, earlier known as Probabilistic Relational Neighbor (pRN) is a simple method that takes advantage of the underlying structure between related elements. It is a node-centric method, that is, it processes one object o_i at a time taking into account a set of n objects in its neighborhood N_i. wvRN estimates $p(o_i|N_i)$, the probability that o_i corresponds to a true object occurrence given its neighborhood N_i, as the weighted mean of the class-membership probabilities predicted by the entities in N_i (see Fig. 2a). It is defined as follows:

$$wvRN(o_i|N_i) = \frac{1}{z} \sum_{o_j \in N_i} p(o_i|r_{ij}).w_j \qquad (1)$$

with z a normalization factor and w_j taking into account the noise in the object detector (see below). $wvRN(o_i|N_i)$ is the relational score of object o_i given its neighborhood N_i. The conditional $p(o_i|r_{ij})$ represents the probability of object o_i occurring given its relation r_{ij} with object o_j. Using Bayes' Rule we estimate $p(o_i|r_{ij})$ as the posterior:

$$p(o_i|r_{ij}) = \frac{p(r_{ij}|o_i)p(o_i)}{p(r_{ij}|o_i)p(o_i) + p(r_{ij}|\neg o_i)p(\neg o_i)} \qquad (2)$$

The components of Eq.2 are obtained through the following procedure. First, we run the local detector on a training set with annotated objects producing a set of hypotheses per image. Then we label the hypotheses as true positives (TP) or false positives (FP) based on the Pascal VOC [6] matching criterion. We define pairwise relations r_{ij} between the hypotheses reported for each image. Relations are divided in two groups. One group contains relations in which both participants are TP hypotheses and the second group contains relations in which at least one participant is a FP hypothesis. Finally, the relations of these groups are used via Kernel Density Estimation (KDE) to estimate $p(r_{ij}|o_i)$ and $p(r_{ij}|\neg o_i)$ respectively. This method captures the statistics of typical configurations. The priors $p(o)$ and $p(\neg o)$ of the object occurring or not at the given location, are estimated as the percentage of TP hypotheses and FP hypotheses in the training set, respectively.

The weighting factor w_j of equation 1 takes into account the noise that is introduced by the object detector in the neighboring objects o_j. We estimate w_j using a *Probabilistic Local Classifier* that takes into account the score s_j provided by the object detector for its respective hypothesis o_j. The output of this classifier will be the posterior $p(o_j|s_j)$ of the occurrence of the object o_j given its score s_j. We compute this posterior following the procedure presented in [23]:

$$w_j = p(o_j|s_j) = \frac{p(s_j|o_j)p(o_j)}{p(s_j|o_j)p(o_j) + p(s_j|\neg o_j)p(\neg o_j)} \qquad (3)$$

The components of this equation are obtained following a procedure similar to that for Eq.2 up to the point where hypotheses are labeled as TPs or FPs. Then, based on the TP and FP hypotheses we compute the conditionals $p(s|o)$ and $p(s|\neg o)$ respectively via KDE. Finally, the priors $p(o)$ and $p(\neg o)$ are estimated in the same way as in Eq.2. As a result, $p(o_j|s_j)$ expresses the probability of a hypothesis be-

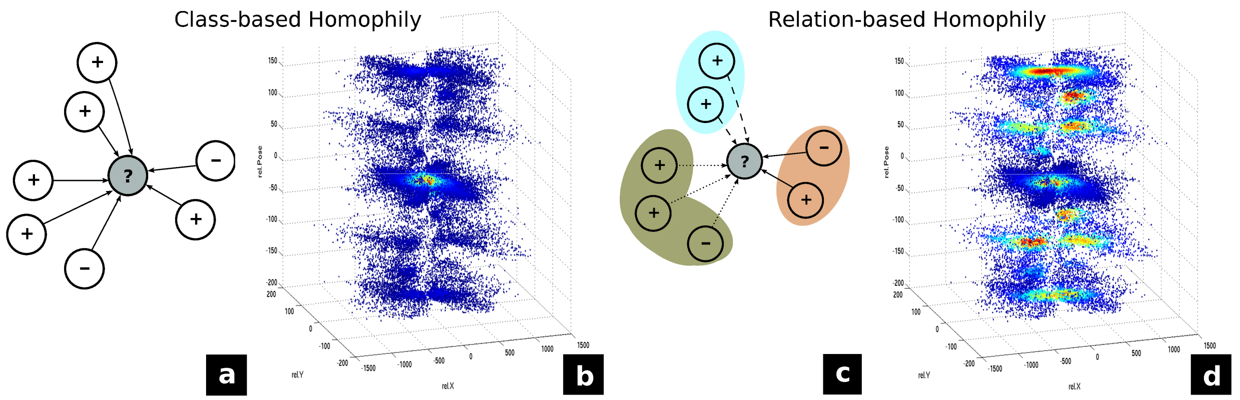

Figure 2. Cautious definition of Homophily. Class-based homophily: a) voting, b) density distribution; and Relation-based homophily c) voting, d) density distribution. Density distributions from RF1 relations on the KITTI dataset. See Sec.5 and Sec.6)

ing correct given its detection score. This procedure allows us to plug-in any standard object detector in our method.

The normalization term z of Eq.1 is estimated as $z = \sum w_j = \sum p(o_j|s_j)$. Finally, the condition $o_j \in N_i$ of the sum in Eq. 1 is of relevance for wvRN inference since wvRN estimates class-membership probabilities based on two assumptions: First, the class of an object depends on its neighbors, and second, the entities exhibit *homophily* in their behavior, i.e. they tend to associate with other objects of the same class. Next, we will specify how to integrate "cautious" inference in this model and propose an alternative type of homophily.

3.2. Cautious Inference

An algorithm is considered "cautious" if it seeks to identify and employ the more certain or reliable relational information [17]. We focus on two factors that [17] introduces to control the degree of caution in an algorithm. The first factor dictates to use only objects for which the prediction is confident enough. The second factor increases caution by favoring already-known relations. These are relations between objects that have been seen in the training images.

For the aggressive version of our relational classifier, we use wvRN as described in Eq. 1. For each object hypothesis, it considers *all* the other objects o_j in its neighborhood N_i during the inference. For the *cautious* version of our relational classifier, we enforce the above principles in the following fashion. For the first principle, giving relevance to certain objects, we perform an iterative approach inspired by [21]. Given a set of hypotheses $O = \{o_1, o_2, ..., o_n\}$, we define the disjoint sets O^k and O^u as the known and unknown objects, respectively, with $O = O^k \cup O^u$ at all times. We initialize $O^k = \{\}$ and $O^u = O$ and flag as *known* object, the hypothesis with the highest score based on the probabilistic local classifier (Eq. 3) . This hypothesis is moved to the set of known objects O^k. Then, the wvRN score for the unknown objects o_i^u is re-estimated considering *only* the known objects o_j^k in their neighborhood N_i. This re-defines Eq. 1 in the following way:

$$wvRN(o_i^u|N_i) = \frac{1}{z} \sum_{o_j^k \in (N_i \cap O^k)} p(o_i^u|r_{ij}).w_j \qquad (4)$$

We flag the hypothesis with highest wvRN score as *known* and move it to the set of known objects O^k. We repeat this procedure promoting one hypothesis o_i^u at a time until the set of unknown objects O^u is empty. Finally, for the sake of similarity in the ranking of the new scores, we re-estimate the score of the first promoted object using Eq. 4 with the second promoted object as known neighbor.

For the second principle of *cautious* inference, favoring relations already seen on training data, our use of KDE for estimating the vote $p(o_i^u|r_{ij})$ for each neighbor object o_j^k implicitly introduces this characteristic in the inference.

Relation-based Homophily: We add and explore an alternative principle to define homophily to use present relations. Our additional principle emphasizes the homophily exhibited in the behavior of the objects. We assume that the pairwise relations estimated from object pairs belong to a set of underlying relationships $Q = \{q_1, q_2, ..., q_m\}$ and that objects tend to associate with each other based on these relationships. As a consequence, homophily will shift from being defined by the object class (as in Eq. 4) to being defined by these relationships. To take into account this principle we compute the neighbor vote $p(o_i^u|r_{ij})$ of Eq. 4 using an intermediate clustering step. First, we extract all the pairwise relations between object hypotheses from a training set of images. We label them as we did for Eq. 2 and run an unsupervised clustering algorithm on them, producing a set of clusters. The centers of these clusters then represent the relationships $\{q_1, q_2, ..., q_m\}$. At test time, each of the relations r_{ij} is assigned to its closest relationship q_{ij} and the TPs and FPs relations found on the training set for the cluster q_{ij} are used to perform KDE. In summary, as Fig. 2c presents, all the known neighboring nodes participate as in Eq. 4 but their vote depends on the relationship that links them. Additionally, comparing the relation-based density distribution (Fig. 2d) with its class-based equivalent (Fig. 2b), one can see that considering underlying relationships has the effect of removing the bias towards the most frequent pairwise relation that is introduced when all the pairwise relations are used for inference (Fig. 2b).

4. Combining Information Cues

At this point, we have two methods to estimate the probability of the occurrence of an object hypothesis o_i: the local classifier, based on appearance, as evaluated by the object detector, and the relational classifier, based on its neighborhood N_i. The reader should note that while the local classifier pulls the decision towards individual features, the relational classifier (Eq.4) pulls it towards the collective feature of group fitting. Given this opposite behavior of our classifiers, local and relational, we need a method to combine them. We follow a method similar to [23]. We use a validation set of images on which we run the object detector. After defining pairwise relations between object hypotheses, we label them as TP and FP hypotheses using the annotations. Then, for each object hypothesis, we compute the score pair (s_{lc},s_{rc}) of the local and relational classifier for each image. For the local classifier, we use the output of Eq.3, applied on o. For the Relational Classifier we use the response of Eq.4. Using these pairs we estimate the conditionals $p(s_{lc}, s_{rc}|o)$ and $p(s_{lc}, s_{rc}|\neg o)$ via Kernel Density Estimation. Finally, the probabilistic score with enforced consistency is estimated as the posterior $p(o|s_{lc}, s_{rc}) = \frac{p(s_{lc},s_{rc}|o)p(o)}{p(s_{lc},s_{rc}|o)p(o)+p(s_{lc},s_{rc}|\neg o)p(\neg o)}$ using Bayes' Rule with $p(o)$ and $p(\neg o)$ determined as for Eq. 2.

5. Implementation Details

This paper studies the impact of *cautious inference*, when reasoning about object relations, for object detection. For this reason rather than proposing our own object detector we use a state-of-the-art detector to acquire evidence of objects in the scene. We build on top of the detector proposed in [14][1] which is based on the popular deformable parts model of [7]; it is designed to jointly tackle the problems of object detection and pose estimation. We use it as an off-the-shelf detector. This detector feeds our framework with confidence scores, locations (2D bounding box) and poses of object hypotheses discretized into 8 partitions.

We define relations between objects in three formats. The first format (RF1) considers differences in x- and y-coordinates $(\Delta x_{ij}, \Delta y_{ij})$ in the 2D image space and the relative pose $\Delta\theta_{ij}$ of the pose θ predicted by the object detector producing a triplet $r_{ij}^{(RF1)} = (\Delta x_{ij}, \Delta y_{ij}, \Delta\theta_{ij})$. The second format (RF2) is based on [13]. In this work relations are represented as a tuple $r_{ij}^{(RF2)} = (rx_{ij}, ry_{ij}, r\rho_{ij}, ra_{ij})$ where $rx_{ij} = x_i - x_j\frac{\rho_i}{\rho_j}$ and $ry_{ij} = y_i - y_j\frac{\rho_i}{\rho_j}$. The factor $\frac{\rho_i}{\rho_j}$ normalizes the translation by object size and is used as a proxy for handling the global scale of the scene. $r\rho_{ij} = \frac{\rho_i}{\rho_j}$ denotes the relative scale ρ_i (the scale of object o_i) and is computed as the square root of the bounding box area of the object. Finally, $ra_{ij} = \frac{a_i}{a_j}$ represents the relative viewpoint, where the viewpoint a_i is encoded by the aspect ratio of the bounding box. The third format (RF3), is purely spatial and considers differences in x- and y-coordinates $(\Delta x_{ij}, \Delta y_{ij})$ only in the 2D image space. This is used in cases where object pose annotations are not available.

In our experiments relationships are discovered using the XMeans [4] clustering algorithm. XMeans is an iterative version of an accelerated KMeans in which the user only provides the range in which K may be located. We provide the range $[4, 64]$ for K to the XMeans algorithm.

Kernel Density Estimation (KDE), with $f(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h})$ is performed using publicly available code[2]. We use a gaussian kernel K, x_i represents each of the n observations (detection score or pairwise relations) gathered from the annotated images, and h is the bandwidth value. This h value is obtained in a data-driven fashion using Silverman's Rule of Thumb [27] , $h = 1.06\hat{\sigma}n^{-1/5}$, where $\hat{\sigma} = min(std(x), iqr(x))$. Kernel products are used for the case of Multivariate KDE.

6. Evaluation

Datasets: We run experiments in the object detection set of the *KITTI benchmark* [10]. We focus on urban scenes with *car* as the class of interest. This dataset contains multiple cars occurring in each image. This provides a challenging realistic scenario with occlusions and clutter that is useful to evaluate the performance of our relational classifier. We evaluate against all the object annotations independent of their occlusion level. We define three sets from the training set of the dataset [10]. First, we divide the sequences that are part of the training set in two sets using the time labels. The images from the first half of all the sequences are used for training while the rest are used for testing, producing two sets with no overlap. Furthermore, the training set is split in two sets for training and validation purposes, producing a total of three sets. In our experiments, the training set is used for extracting the pairwise relations used to perform KDE in the relational classifier and to discover relationships. The validation set is used for learning the combination of the local and the relational classifier. This dataset was obtained using a car-mounted camera and resembles the settings used for autonomous navigation. Additionally, we run experiments on the *MIT-StreetScenes* (MITSS) dataset [1]. Different from the KITTI benchmark, this dataset was obtained using a

[1] http://agamenon.tsc.uah.es/Personales/rlopez/data/pose-estimation/

[2] http://www.ics.uci.edu/ ihler/code/kde.html

Dataset **KITTI benchmark**	**Relations Representation : RF1**								**Relations Representation : RF2**							
	Class-based Homophily				Relation-based Homophily				Class-based Homophily				Relation-based Homophily			
	Global		Near		Global		Near		Global		Near		Global		Near	
Set	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.
2-3	0.33	0.35	0.32	0.33	0.33	0.35	0.27	0.34	0.32	0.32	0.36	0.34	0.40	0.39	0.37	0.36
4-7	0.31	0.40	0.30	0.29	0.30	0.40	0.28	0.35	0.34	0.43	0.38	0.34	0.44	0.53	0.41	0.49
8+	0.28	0.36	0.27	0.23	0.26	0.36	0.26	0.30	0.30	0.39	0.37	0.29	0.40	0.51	0.40	0.44
all	0.29	***0.38***	0.29	0.26	0.28	***0.37***	0.27	0.32	0.32	0.40	0.37	0.31	0.41	***0.50***	0.40	0.45

Dataset **MIT StreetScenes**	**Relations Representation : RF3**								**Relations Representation : RF2**							
	Class-based Homophily				Relation-based Homophily				Class-based Homophily				Relation-based Homophily			
	Global		Near		Global		Near		Global		Near		Global		Near	
Set	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.	aggre.	caut.
2-3	0.71	0.74	0.65	0.58	0.70	0.71	0.62	0.60	0.70	0.69	0.68	0.61	0.67	0.68	0.68	0.66
4-7	0.54	0.65	0.46	0.42	0.49	0.59	0.48	0.50	0.49	0.56	0.50	0.48	0.47	0.55	0.52	0.57
8+	0.35	0.46	0.30	0.29	0.33	0.45	0.29	0.39	0.33	0.42	0.38	0.34	0.33	0.43	0.39	0.47
all	0.54	***0.63***	0.47	0.43	0.51	0.59	0.47	0.50	0.51	***0.56***	0.52	0.48	0.49	0.55	0.53	***0.57***

Table 1. Mean Average Precision of the Relational Classifier for object detection on the KITTI and MITSS datasets. (Only using context to predict the object presence)

consumer camera and offers more viewpoint variability. For our experiments we divided this dataset in 4 subsets. The first two quarters were used for training and validation while the third and fourth quarters were used for testing. We run experiments on three splits of each dataset and report mean performance results. In addition, in order to check the behavior of the object detector, the relational classifier and the combination of the two, we split the test set in three subsets. These subsets are defined based on the number of hypotheses available for the inference stage. The subsets contain images with [2,3],[4,7] and $[8,\infty)$ hypotheses respectively.

Experiment: We reason about object relations as means for object verification, i.e. to correct errors of the object detector. We define *Object Verification* as the task of reranking the set of hypotheses given by a detector in such a way that the most likely hypotheses get a higher score. The task of object verification is evaluated procedure used in Pascal VOC [6]. We report results using mean Average Precision (mAP) as performance metric.

We report experiments with eight baselines defined by the combination of three parameters. The first parameter *Neighborhood Scope*, indicates how the neighborhood N_i of a particular object o_i is defined based on their relative location. It is set to "global" if it considers *all* the objects despite their location. It is set to "near" if it only considers the objects within a relative distance t where t is defined as the median distance of all the spatial relations in the training set. This parameter represents the Markovian assumption that some relational methods enforce by only considering neighboring objects in their spatial vicinity. The second parameter indicates the type of inference to use which can be "aggressive" (Eq. 1) or "cautious" (Eq. 4). The last parameter *Homophily drive* covers the possible causes that relate entities. It can be driven by the *class* of the object, as in traditional homophily, or by the *relationships* that we propose in this work. We present results using the relation representations RF1 and RF2 (Sec. 5) for the KITTI dataset. For the case of MITSS we use representations RF2 and replace RF1 with RF3 (Sec. 5) due to the lack of annotated object poses. Table 1 shows the performance of the different baselines when only the relational classifier is used, that is, only considering contextual information. Table 2 shows the performance of the combination of local and relational classifiers for the top performing baselines. Note that the baseline defined by *aggressive* inference with RF3 relations, assuming *Class-based Homophily* in a *Global* Neighborhood, is a *Things*-based version of [23].

Dataset **KITTI benchmark**		**RF1**		**RF2**	
		Class-based Homophily		Relation-based Homophily	
		Global		Global	
Set	Detector [14]	aggre.	caut.	aggre.	caut.
2-3	0.65±0.027	0.65±0.022	0.66±0.020	0.66±0.033	0.64±0.017
4-7	0.63±0.010	0.64±0.009	0.66±0.016	0.67±0.017	0.71±0.019
8+	0.60±0.011	0.59±0.007	0.61±0.004	0.63±0.004	0.68±0.009
all	0.61±0.011	0.61±0.009	0.63±0.007	0.65±0.011	***0.68±0.003***

Dataset **MIT StreetScenes**		**RF3**		**RF2**	
		Class-based Homophily		Class-based Homophily	
		Global		Global	
Set	Detector [14]	aggre.	caut.	aggre.	caut.
2-3	0.74±0.005	0.83±0.007	0.86±0.002	0.79±0.009	0.80±0.011
4-7	0.68±0.005	0.77±0.001	0.81±0.031	0.73±0.004	0.77±0.016
8+	0.68±0.033	0.69±0.003	0.71±0.044	0.68±0.043	0.70±0.030
all	0.69±0.006	0.77±0.001	***0.80±0.028***	0.73±0.011	0.76±0.014

Table 2. Mean Average Precision of the top performing baselines of the combination of Local [14] and Relational Classifiers for object detection on the KITTI and MITSS datasets. Note that the baseline defined by *aggressive* inference with RF3 relations, assuming *Class-based Homophily* in a *Global* Neighborhood, is a *Things*-based version of [23].

Dataset **KITTI benchmark**		**RF3**		**RF2**	
		Class-based Homophily		Relation-based Homophily	
		Global		Global	
Set	Detector [7]	aggre.	caut.	aggre.	caut.
all	0.65±0.003	0.68±0.007	0.71±0.007	0.72±0.009	**0.75±0.003**

Dataset **MIT StreetScenes**		**RF3**		**RF2**	
		Class-based Homophily		Class-based Homophily	
		Global		Global	
Set	Detector [7]	aggre.	caut.	aggre.	caut.
all	0.62±0.004	0.66±0.011	**0.71±0.012**	0.65±0.026	0.69±0.014

Table 3. Mean Average Precision of the top performing baselines of the combination of Local [7] and Relational Classifiers for object detection on the KITTI and MITSS datasets. Note that the baseline defined by *aggressive* inference with RF3 relations, assuming *Class-based Homophily* in a *Global* Neighborhood, is a *Things*-based version of [23].

Discussion: Overall, based on the parameters previously mentioned, the performance of the evaluated algorithms present the following trend: First, and maybe somewhat surprisingly, on average, *global* neighborhoods provide higher performance than the *near* option. Second, on the scope of a *global* neighborhood, *cautious* methods outperform their *aggressive* counterparts. Third, dataset-wise, *Relation-based Homophily* performs better in the KITTI dataset, where camera settings are more constrained. This may suggest that the method to uncover relationships may be sensible to changes in viewpoint. Finally, the proposed *cautious* scheme boosts the performance of the baselines [7, 14, 23]. Now we discuss the results in more detail.

Regarding the *relations format*, the difference in performance of RF2 on the different datasets in Table 1 suggests that RF2 is better suited for working on constrained camera settings, as in the KITTI dataset. Furthermore, the difference in performance between RF1 and RF3, shows a weakness of relational methods when relations are defined from, possibly, unstable attributes. In this case, the relative pose information used in RF1 may be the cause of its relatively lower performance.

Regarding the *type of inference* to use, both Tables 1 and 2, show that *cautious* reasoning with object relations always outperforms its aggressive counterpart when exercised on a *global* neighborhood. This is supported by mean improvements, over traditional *aggressive* inference, of 8%, on the Relational Classifiers (Table 1), and 2.5% on the combination of Local and Relational classifiers (Table2). In addition, there is an improvement of 5% and 3% over the baselines [14] and [23], respectively.

Related to the alternative notion of *homophily*, Relation-based homophily outperforms class-based homophily on a *global* neighborhood when using RF2. This is opposite to what is seen with the related RF1 and RF3 where class-based homophily performs better. It seems that, similar to RF2, Relation-based homophily performs better in constrained settings, with lower viewpoint variability as in KITTI. In this context, the representation used for the relations plays a relevant role since the clustering method used to discover the underlying relationships operates directly on the attributes of the pairwise relations. Likewise, the method to discover these underlying relations affects the inference process, i.e. boundary effects that can be introduced by hard clustering methods as the one employed in this work. Future work will focus on analyzing the influence that the selected method for discovering the relationships has on relation-based homophily. The mean boost in performance of 8.5% on the relational classifier makes relation-based homophily an appropriate principle in scenarios where no local information is available on the unknown object. Indeed, it is remarkable that the cautious relational classifiers, only using context information, can get as low as 8% behind the local detector for their top performing cases. Note in Fig. 3 how the baselines based on cautious inference effectively promote hypotheses that had been ranked low by the detector.

The change in performance obtained by the local classifier, the object detector, and the relational classifier in the different subsets of images hints at the scenarios for which each classifier is better. For the local classifier, its performance is at its highest point when a low number of hypotheses is reported and decreases as the number of hypotheses increases. This represents the scenario with few, possibly non-overlapping, hypotheses (see Fig. 3 top row). On the other hand, the relational classifier performs better as the number of hypotheses increases (see Fig. 3 and Fig. 4). This proves their "competitive" behavior. For the combination of the two classifiers, there is a peak in performance in the second subset of images. It should be noted that the following subset, where performance drops, is the one with higher number of hypotheses, thus, more likely to contain a larger proportion of false hypotheses. This suggests that the combination of classifiers may be weak towards large occurrence of false hypotheses. This can be supported by the fact that there is a true positive - false positive ratio of 0.24 and 0.28 for the [14] baseline on the KITTI and MITSS datasets, respectively. We see that the combination of the responses of the local and relational classifiers produces an average and maximum improvement of 5% and 9%, respectively over the [14] baseline.

For object detection the use of a neighborhood with reduced spatial scope is discouraged since it has relatively lower improvement of 1.3% than when reasoning in *global* neighborhoods where a mean improvement of 4.7% was obtained over different relations representations.

Finally, we ran experiments using [7] to generate the initial object hypotheses and defined RF2 and RF3 relations between objects. Table 3 shows how results follow a similar trend as the ones obtained in the other experiments.

7. Conclusions

We showed that *cautious* inference about object relations outperforms traditional *aggressive* inference methods for object detection. *Cautious* methods empirically provided mean improvements of 8% and 2.5% on relational and combined classifiers, respectively, over its *aggressive* counterparts. Furthermore, we have introduced a notion of *relational-homophily* that recovers underlying structures from the observed relations aiming to better understand the behavior of the related classes of interest and improve inference. Improvements of 8.5% on purely relational methods makes *relational-homophily* a promising principle to use when local information about the unknown instances is not available (e.g. in an inpainting scenario). Furthermore, experiments suggest that performing cautious inference paired with Relation-based homophily with relations in RF2 representation is beneficial for more camera constrained settings such as the ones found in systems for autonomous navigation. Future work will focus on three directions: better representations for reasoning in 3D space, which typically outperform methods that operate in 2D; better methods to

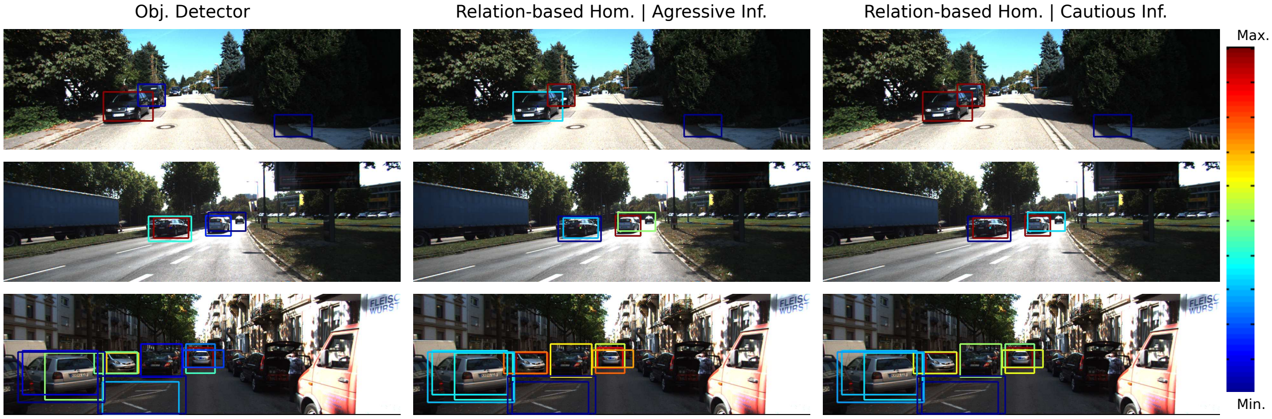
Obj. Detector
Relation-based Hom. | Agressive Inf.
Relation-based Hom. | Cautious Inf.
Max.
Min.

Figure 3. Qualitative results in a Global Neighborhood setting. Confidence scores color coded in jet scale. Note how Cautious Inference promotes hypotheses with initial low score. (Best viewed in color)

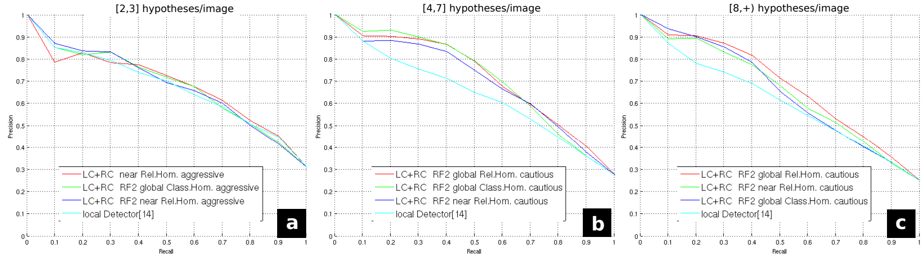

Figure 4. Precision-Recall curves for the top 3 ranking baselines on the KITTI dataset for the different image sub-sets based on their respective number of hypotheses: a) [2,3], b) [4,7] , and c) [8,∞) hypotheses/image respectively.

recover the underlying structures of the relational space defined by the object relations and investigating the generality of these observations in the context of other object categories or other application scenarios .

Acknoledgments: This work is supported by the ERC grant 240530 COGNIMUND, the VASI Project and the BOF Research Fund KUL 3E100864.

References

[1] S. M. Bileschi. Streetscenes: Towards scene understanding in still images. In *PhD Dissertation, MIT*, 2006. 5

[2] M. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010. 2

[3] R. G. Cinbis and S. Sclaroff. Contextual object detection using set-based classification. In *ECCV*, 2012. 2

[4] A. M. Dan Pelleg. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, 2000. 5

[5] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 2011. 1, 2

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. 3, 6

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *TPAMI*, 2010. 2, 5, 6, 7

[8] D. A. Forsyth, J. Malik, M. M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler. Finding pictures of objects in large collections of images. In *ECCV*, 1996. 2

[9] C. Galleguillos, B. McFee, S. Belongie, and G. R. G. Lanckriet. Multi-class object localization by combining local contextual interactions. In *CVPR*, 2010. 1

[10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 5

[11] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008. 1, 2

[12] A. Jain, A. Gupta, and L. S. Davis. Learning what and how of contextual models for scene labeling. In *ECCV*, 2010. 1, 2

[13] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for scene understanding. In *CVPR*, 2012. 1, 3, 5

[14] R. J. Lopez-Sastre, T. Tuytelaars, and S. Savarese. Deformable part models revisited: A performance evaluation for object category pose estimation. In *Workshop at ICCV*, 2011. 5, 6, 7

[15] S. A. Macskassy and F. Provost. A simple relational classifier. In *MRDM*, 2003. 2

[16] S. A. Macskassy and F. J. Provost. Classification in networked data: A toolkit and a univariate case study. *JMLR*, 2007. 1, 2, 3

[17] L. McDowell, K. M. Gupta, and D. W. Aha. Cautious inference in collective classification. In *AAAI*, 2007. 1, 3, 4

[18] L. McDowell, K. M. Gupta, and D. W. Aha. Cautious collective classification. *JMLR*, 2009. 1, 2

[19] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 2001. 1

[20] J. Neville and D. Jensen. Leveraging relational autocorrelation with latent group models. In *ICDM*, 2005. 2

[21] J. Neville and D. D. Jensen. Iterative classification in relational data. In *Workshop on SRL at AAAI*, 2000. 2, 4

[22] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001. 2

[23] R. Perko and A. Leonardis. A framework for visual-context-aware object detection in still images. *CVIU*, 2010. 1, 2, 3, 5, 6, 7

[24] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 1, 3

[25] P. Sen, G. Namata, M. Bilgic, and L. Getoor. Collective classification. In *Encyclopedia of Machine Learning*. 2010. 1, 3

[26] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *CVPR*, 2011. 2

[27] M. Wand and M. Jones. Kernel smoothing, 1995. Chapman & Hall CRC. 5