Relationship Matters: Relation Guided Knowledge Transfer for Incremental Learning of Object Detectors

Kandan Ramakrishnan[1,2], Rameswar Panda[1,2], Quanfu Fan[1,2],
John Henning[1,2], Aude Oliva[2,3], Rogerio Feris[1,2]
[1]IBM Research, [2]MIT-IBM Watson AI Lab, [3]MIT CSAIL

Abstract

Standard deep learning based object detectors suffer from catastrophic forgetting, which results in performance degradation on old classes as new classes are incrementally added. There has been a few recent methods that attempt to address this problem by minimizing the discrepancy between individual object proposal responses for old classes from the original and the updated networks. Different from these methods, we introduce a novel approach that not only focuses on what knowledge to transfer but also how to effectively transfer for minimizing the effect of catastrophic forgetting in incremental learning of object detectors. Towards this, we first propose a proposal selection mechanism using ground truth objects from the new classes and then a relation guided transfer loss function that aims to preserve the relations of selected proposals between the base network and the new network trained on additional classes. Experiments on three standard datasets demonstrate the efficacy of our proposed approach over state-of-the-art methods.

1. Introduction

Deep Convolutional Neural Network (CNN) based object detectors [9] have achieved state-of-the-art results on datasets such as PASCAL VOC [7]. While they perform very well on standard benchmark datasets, a challenge for real world applications is learning object detectors incrementally, where new classes are added over multiple training sessions. One strategy for training models with new classes is to fine-tune networks on the new data. While the performance of a finetuned network on the new classes is satisfactory, its performance on old classes degrades significantly [20]. This is a well known problem that occurs when training CNNs called catastrophic forgetting [26].

A number of recent works have addressed the issue of catastrophic forgetting in CNNs. Most of the work is either model-based [1, 3, 17] or data-based [20, 6, 37]. In model based approaches, existing works look at different methods

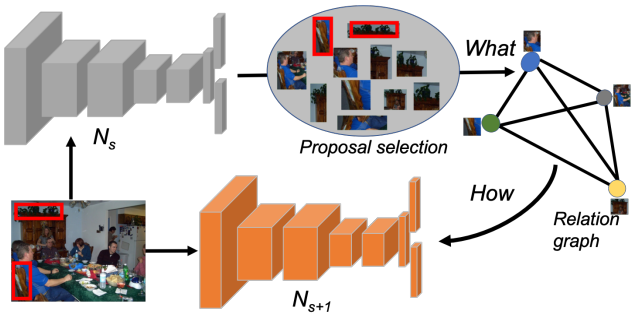

Figure 1. **Overall framework** of our proposed approach. Incremental learning of network N_{s+1} trained from a base network N_s which consists of i) proposal selection for *what* knowledge to transfer and ii) relation transfer for *how* knowledge is used for guiding the network training for incremental object detection.

to preserve important weights of the network for the previously learnt categories and use the remaining neurons for learning new classes [17]. On the other hand, data-based approaches mainly use some measure of the data representation in the network, for example knowledge distillation [12] and its variants. Distillation loss [12] is mainly adapted to maintain the responses of the network on the old tasks whilst updating it with data for the new training classes [20]. Alternatively, few works [19, 2] also store some part of the old data which can be used during the new training to partly alleviate the problem of catastrophic forgetting.

Most of the works in incremental learning focus on image classification, and not much work has been done in the domain of object detection. Incremental object detection is more challenging [37] since it has an additional challenge of localizing the objects. Furthermore, in image classification old and new class examples are usually distinct while in object detection it is likely that both old and new classes co-occur in the same image. Thus during training, the network has to learn to detect old classes that might be present in new data. A recent work that tackles incremental learning of object detectors [37] uses a variant of knowledge distillation on a selected set of object proposals that guide the learn-

ing of new categories over a base model that is previously trained on a set of categories. However the object proposals which guide the learning process (i.e *what knowledge* to be transferred from old to the new network), are randomly selected which severely affects the overall performance the newly learned object detector.

Objects in images show natural relations, for example a television is more likely to be present in front of the couch and less likely to be present in an image containing horse. Such object relations can serve as useful priors for object detection. Thus, using not only the representation of objects but also the relations that objects exhibit is a more comprehensive way of showing the model *how to transfer knowledge*. Motivated by this, we propose a relation-based loss to transfer knowledge from base network to new network. Specifically, we introduce a similarity preserving knowledge transfer loss that guides the training of the new network such that the relationship between objects are preserved in the new network. To effectively transfer relations, we additionally introduce a proposal selection mechanism by exploiting the ground truth of new classes as well as the high confidence proposals from the old network. The overall approach is illustrated in Figure 1.

Our approach works as follows. We first use pre-computed proposals from Edge Boxes [46] to train a Fast-RCNN object detector [9] on a set of base classes. Given a set of new classes, the incremental learning phase transfers knowledge from the base model that was trained on the old set of classes to the new model while also learning the new classes. A subset of object proposals are selected from the pre-computed proposals such that these proposals capture the most informative regions in the image related to the classes. This selection mechanism ensures that proposals related to both new and old classes are used for determining the relations between proposals. Based on the selected proposals, a relation matrix is computed using the Euclidean distance and the associated loss is used to penalize the divergence between base and new network.

We evaluate our proposed approach on multiple benchmark datasets [7, 22] to show the benefit of relation-guided knowledge transfer for incremental learning of object detectors. Additionally we evaluate our approach with multiple episodic training of the network, i.e, multiple instances of new classes being added to the existing set of classes. Our results show that the performance of base classes are preserved effectively even in the case of multiple episodes by using our proposed approach as compared to the baselines.

Our main contributions can be summarized as follows.

- We use a *proposal selection* mechanism that utilizes ground truth as priors for selecting *what knowledge* to transfer in incremental object detection.
- We introduce a novel *relation guided transfer loss* for *how knowledge* from the old network is used to supervise the training of the new network.
- We show competitive performance on three datasets and use different measures that present a fair way to evaluate models for incremental object detection.

2. Related Work

Our work relates to three major research directions: object detection, incremental learning and knowledge distillation. Here, we focus on some representative methods closely related to our work.

2.1. Object Detection

There is a continued interest in the vision community on learning object detection models using deep CNNs [10, 9, 11] as they significantly outperform traditional methods [8, 33, 41]. Generally, modern CNN-based object detection frameworks fall into two groups. One is the two-stage detectors like R-CNN [10], Fast R-CNN [9], Faster R-CNN [32], Deformable CNN [5], Mask R-CNN [11], etc. The second one is one-stage detectors such as OverFeat [35], YOLO [31], SSD [23] and RetinaNet [21], etc which have also been proposed driven by the requirement of real time inference in many applications. Anchor free detectors have also been proposed that use keypoint estimation for efficient object detection [18, 45]. However, all of these methods focus on learning detectors with a fixed set of classes unlike the problem domain we consider where the number of classes keeps growing. Incremental learning of object detectors still remains as a novel and largely under-addressed problem in computer vision.

2.2. Incremental Learning

Incremental (a.k.a continual) learning has been studied from multiple perspectives (see [27] for a recent survey). Broadly speaking, the existing works can be divided into two main types: one is task-incremental learning where the number of tasks (i.e., datasets) keeps growing [1, 3, 17, 20, 13]. Another is class-incremental learning where the number of class labels keeps growing [30, 6, 19, 2]. The class-incremental problem is more difficult than the task-incremental problem as the model can often confuse the new class with a base class [3]. Various strategies have been studied for both incremental learning scenarios including model-based approaches that constraint the network updates to be around the original values [1, 3, 17], data-based approaches [6, 20] that keep the knowledge of the previous tasks by knowledge distillation [12] and a combination of both for the better performance [16]. Memory-based approaches [25, 14] or generative models [36] have also been proposed for incremental learning. However, all these approaches consider incremental learning of image classifiers

unlike the problem domain we consider. Specifically, we focus on the more challenging incremental object detection task, where it is very common for the old and the new classes to co-occur, unlike the classification task.

The most relevant work to ours is the incremental object detection work (IOD-KD) proposed in [37] that uses distillation loss [12] to preserve the knowledge of base classes without storing the data of base classes. However, our approach and the work IOD-KD in [37] have significant differences. First, the object proposals (*what*) selected for transferring knowledge in IOD-KD are chosen at random. In contrast, our approach selects proposals that overlap with ground truth objects by avoiding large amount of noise from unrelated areas. The intuition is that object detectors care more about local regions that overlap with ground truth objects and hence exploiting ground truth bounding boxes as priors for selecting what knowledge to transfer can be helpful in incremental object detection. Second, we intend to use object proposal relations in an incremental setup, instead of only knowledge distillation [12], to transfer more comprehensive knowledge of base classes from the old model to the new model (*How*). We hypothesize that proposal relationships encode a detector's representation more precisely. Hence, constraining the proposal relationship using a relation guided transfer loss, to minimize the divergence of the representations of new detector from that of an old one is more meaningful. A very recent work [44] addresses the problem of incremental object detection using deep model consolidation with auxiliary data which are harder to obtain in many cases and often not feasible when the memory budget is limited.

2.3. Knowledge Distillation

Knowledge distillation [12] that focuses on transferring knowledge from a large network to a small one for model compression has attracted intense attention in the recent years. Much progress has been made in developing a variety of ways through matching logits [12], intermediate features [34], attention maps [43], and feature space transformation [42]. Leveraging feature similarity [24, 28, 38] is also another recent trend for knowledge distillation. With growing interests in knowledge distillation, task-specific KD methods have also been proposed for object detection [39, 4, 40, 29] but with fixed number of classes. Compared to all these works, in this paper, we focus on the more difficult problem of learning object detectors incrementally where selecting the right knowledge (*what*) is also equally important including the transfer approach.

3. Proposed Method

Our relation guided knowledge transfer approach for incremental learning of object detectors is shown in Figure 2. Given a network that is trained with the old set of classes,

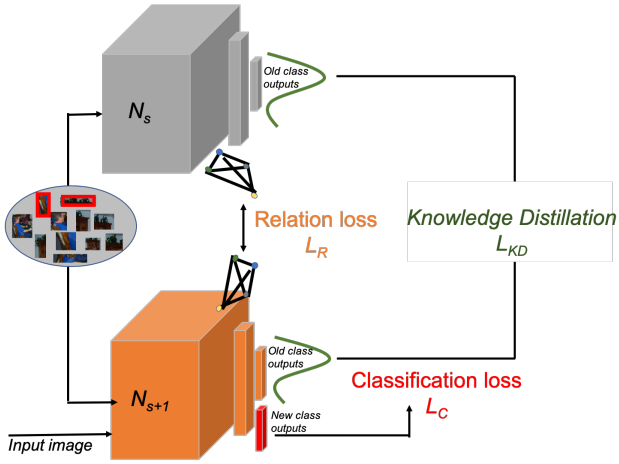

Figure 2. **Relation Guided Knowledge Transfer**: Illustration of our relation guided knowledge transfer approach that combines relation loss with knowledge distillation and classification loss for incremental learning of object detectors.

the incremental learning problem is to train the network with the new set of classes. Let N_s be the base network trained at step s with the old classes. At each incremental step $s+1$ the network N_{s+1} is trained for a new set of categories using the base model. We describe our approach which consists of a proposal selection mechanism and a loss function that transfers knowledge from N_s to N_{s+1}.

3.1. Proposal selection

A main component for incrementally learning new classes for object detection, is the selection of proposals that will be used for transferring knowledge from old network to new network. In [37], given a set of proposals the method randomly selects a small subset of proposals based on the proposal score for non-background class. However, the random selection might miss proposals that are useful to preserve the information from the old network, and we need to select proposals that are relevant for both old and new classes to effectively transfer knowledge.

To overcome the shortcomings from random selection, we propose a selection mechanism that uses the ground truth object bounding boxes as well as proposals with the lowest background score as regions of the image for sampling proposals. Given a set of proposals, for the proposal selection mechanism we start with an initial set of proposals P determined by the lowest background scores (by passing all proposals through the old network we obtain a background score). To select proposals for distillation, we first define image regions R that have a high likelihood of objects. The image regions are generated using the top B ($P \gg B$) lowest background score proposals along with ground truth bounding boxes G as to give us a set of re-

Figure 3. **Example images** from VOC 2007 dataset that show how different objects are related to each other. Objects in yellow bounding box are close to each other, while object in red bounding box is not. Our approach aims to select related proposals and preserve object relations. Best viewed in color.

gions $R = B + G$ for sampling proposals. In the next step, we select proposals from the pool P using the image region R that we defined. Since the image regions contain informative objects, we want to select proposals that are near R. To this end, we measure locality of each proposal from the pool in relation to the defined image region. The locality of a proposal to each image region is evaluated using IOU (intersection over union) and normalized l_2 distance. For each proposal P_j (the j^{th} proposal in the pool) the locality L from each region R_i (the i^{th} region) is calculated using the following formula:

$$L_{i,j} = 1 - \frac{l_2(R_i, P_j)}{max(l_2(R_i, P))} + IOU(R_i, P_j) \quad (1)$$

The l_2 distance is the euclidean distance between the feature representation of the proposal P_j to each region R_i. The IOU is the overlap of the proposal bounding boxes. For each proposal, we then take the maximum localities M across all regions:

$$M_j = max(L_{:,j}) \quad (2)$$

We select the final set of proposals based on the highest locality value given by M and then use them in computing the relation loss as described in the following sections.

3.2. Relation Loss

Given a selected set of object proposals, our goal is to learn a new set of classes while preserving the network performance on the old set of classes. Unlike image classification, in object detection we observe that multiple classes can co-occur in the same image and they exhibit natural relations. For example, a chair and person are highly likely to occur in the same image as shown in Figure 3. Our hypothesis is that relationships encode the network representations more precisely for the problem of object detection. In our proposed approach we use a matrix to encode the relations between every pair of object proposals.

In the incremental learning problem, we start with a base model N_s trained on the old set of categories. At each incremental step, we train a network N_{s+1} on the new set of categories. Once the proposals are selected as described in Section 3.1, we extract feature representations $f(p_i)$ for $i = 1, \ldots, P$ for each of the selected P proposals from the base model N_s and new network N_{s+1}. The features are used to compute a relational matrix, A_s for the base network and A_{s+1} for the new network. The euclidean-distance based matrix is computed for N_s as;

$$A_s(i,j) = ||f(p_i) - f(p_j)||_2, \quad i,j = 1,..,P \quad (3)$$

and similarly for N_{s+1} as;

$$A_{s+1}(i,j) = ||f(p_i) - f(p_j)||_2, \quad i,j = 1,..,P \quad (4)$$

The matrices that capture relations between selected object proposals are used to constrain the learning of the new network and better transfer knowledge from the base model to the new network. For this we define a loss, called relation loss (L_R) which is the l_2 norm between two relation matrices. Given the relation matrices, A_s of the base network and A_{s+1} of the new network, the loss L_R is given by;

$$L_R = ||A_s - A_{s+1}||_2 \quad (5)$$

As shown in Figure 2, the relation loss L_R aims to minimize the divergence in pair-wise distances of selected proposals between the old and new network. The proposal relations provides a strong learning signal for training the new network without forgetting the old classes.

3.3. Relation Guided Knowledge Transfer

Our proposed approach denoted as RKT combines the relation loss L_R with distillation L_{KD} and classification loss L_c as shown in Figure 2. For the distillation loss, logits computed for the proposals by N_s serve as targets for the new network N_{s+1}. As in [37], the mean over the class dimension from unnormalized logits f of each RoI is subtracted to obtain the corresponding centered logits (f) used in the distillation loss. Bounding box regression outputs b (of the same set of proposals used for computing the logit loss) also constrain the loss of the network.

$$L_{KD} = \frac{1}{N} \sum [(f_s - f_{s+1})^2 + (b_s - b_{s+1})^2] \quad (6)$$

The classification loss function per ROI to train the Fast R-CNN detector [9], is given by:

$$L_C(p, k^\star, t, t^*) = -log p_k^\star + [k^\star \geq 1] R(t - t^\star) \quad (7)$$

In the above equation, p is the softmax output of the network for all the classes, t is the output of bounding box layer while $t^\star$ is the ground truth bounding box and $k^\star$ is a ground truth class. While the first part of the loss function corresponds to the classification loss and the second part represents the localization loss as in [9].

The total loss function is defined as;

$$L_{RKT} = \lambda_r L_R + \lambda_c L_C + \lambda_d L_{KD} \tag{8}$$

where λ_r, λ_c and λ_d are the hyperparameters that balance the different loss terms.

3.4. Training procedure

We follow the same training procedure and settings as outlined in [37] for our approach. Following [37], we use Fast-RCNN for object detection since we can exploit proposal selection as a mechanism to guide our learning. We use Edge boxes [46] to pre-compute proposals for all the datasets and the network input is an image with 2000 precomputed proposals represented as bounding boxes.

At step s, a Fast-RCNN [9] base network N_s is trained on a set of categories for detection. At step $s+1$, a new network N_{s+1} is trained for the new set of categories. The new network N_{s+1} is a copy of N_s that is adapted for the new classes. The adaptation is done on the last fully connected classification and bounding box regression layers. Fully connected layers are created for new classes only and the outputs concatenated with the original ones. The new layers are initialized randomly in the same way as the corresponding layers in Fast R-CNN.

Once the network is initialized, each image and corresponding proposals serve as input to both the networks. Based on the scores computed for the proposals using network N_s and the ground truths, a subset of proposals are selected as described in Section 3.1. These set of proposals are used for knowledge transfer from base to new network. For knowledge transfer we compute the L_{RKT} loss as outlined in Section 3.2. During inference, the high-scoring proposals are refined according to bounding box regression. Then, a per-category non-maxima suppression (NMS) is performed to get the final detection results.

4. Experiments

In this section, we present extensive experiments to demonstrate the effectiveness of our proposed approach for incrementally learning of object detectors.

4.1. Datasets and Evaluation

We evaluate the performance of our approach using three datasets: (i) PASCAL VOC 2007, (ii) PASCAL VOC12 dataset and iii) KITTI dataset. Both the VOC datasets [7] have 20 object classes. While VOC07 dataset is divided into 2 subsets, such as trainval containing 5011 images and test containing 4952 images, the VOC12 detection benchmark consists of 5717 images for training and 5823 for testing. We use the standard test splits for evaluation on both VOC datasets. For the KITTI dataset [22], we use the split of 3712 images for training and 3769 for testing.

We use the standard mean average precision (mAP) at 0.5 IoU threshold (i.e., a predicted bounding box is correct if its intersection over union with the ground truth bounding box is higher than 0.5) as the evaluation metric. In addition, following [15], we also compute another metric called Ω_B by dividing the performance of an incremental learning method with the performance of an offline base model trained with all the categories, which we assume is the ideal performance. Both of the metric represents the model's ability to retain prior knowledge while still learning new knowledge. While only mAP is used to asses the performance of the model at the end of the last task in [37], we believe that performance of the model at every episode rather than the end of the learning better characterizes the dynamic aspects of incremental learning. Thus, we consider mAP and Ω_B at each episode as the performance measures for a fair evaluation in incremental learning.

4.2. Experimental Settings

We use SGD with a mini-batch size of 2 images to train the networks in all our experiments. While learning the base network, the initial learning rate is set to 0.001 and decreased by a factor of 10 after each 30K iterations. We use a learning rate of 0.0001 while doing incremental learning over the previously learned model. We also use a weight decay parameter of 0.00005. We train the networks for 40 epochs on both datasets. We adopt the same settings followed by [15] to integrate ResNet into the Fast R-CNN. We apply per-class NMS with an IoU threshold of 0.3 and a batch consists of 64 proposals per image, with 16 of them having an IoU of at least 0.5 with a groundtruth object. We also filtered all the proposals to have IoU less than 0.7 [46]. For incremental learning, we follow the same class ordering (alphabetical) as in [15] and select a subset of classes to learn at each episode. We did not use annotations of all the other classes except the ones that are used in the current episode while learning a network in the incremental set up. We use 10 lowest background score proposals from the base network for computing locality in all our experiments.

4.3. Baselines

Our main baseline is IOD-KD [37] that uses randomly selected object proposals and a sum of a classification loss, bounding box regression loss and distillation loss to transfer knowledge from the old to the new model. Our approach utilizes relational knowledge transfer loss (Eq. 8) over the object proposals selected using ground truth priors in addi-

tion to the same classification and regression loss to consistent with the baseline. We additionally compare with elastic weight consolidation (EWC) [17], which is a model-based approach to regularize the parameter updates using Fisher Information while learning the new classes. Furthermore, we compared with another simple baseline approach for addition of new classes through fine-tuning the old network by replacing the last layer (denoted as *Fine-Tune* in our work) without any knowledge transfer. We use the publicly available code for the IOD-KD implementation and set the hyper-parameters as recommended in the published work.

Method	mAP_{1-19}	mAP_{20}	mAP	Ω_B
N_{1-19}	67.8	-	-	-
$+N_{20}$ w/ Fine-tune	25.0	52.1	26.4	0.38
$+N_{20}$ w/ IOD-KD [37]	67.1	58.1	66.7	0.97
$+N_{20}$ w/ RKT (Ours)	67.6	59.4	67.2	0.98
N_{1-20}	68.2	69.3	68.3	1.00

Table 1. **Learning 19+1 Classes in VOC07 Dataset**. Results show the addition of one class i.e., "tvmonitor" class to a pre-trained detection network trained with 19 classes. Our approach RKT outperforms IOD-KD [37] baseline on both measures.

Method	mAP_{1-10}	mAP_{11-20}	mAP	Ω_B
N_{1-10}	65.8	-	-	-
+ N_{11-20} w/ Fine-tune	13.0	61.8	37.4	0.54
+ N_{11-20} w/ IOD-KD [37]	66.6	58.2	62.4	0.91
+ N_{11-20} w/ EWC [17]	31.6	61.0	46.3	0.67
+ N_{11-20} w/ RKT (Ours)	67.1	59.2	63.1	0.92
N_{1-20}	67.9	68.7	68.3	1.00

Table 2. **Learning 10+10 Classes in VOC07 Dataset.** Results show the the addition of 10 classes, all at once, to a pre-trained object detection network trained initially with 10 classes. The proposed approach outperforms all the baselines.

4.4. Results and Analysis

Table 1-8 show the results of our method and other baselines under different incremental learning scenarios on both VOC datasets. We show that we can retain the performance of base classes for a longer time when more classes are incrementally added to the classifier by using our proposed approach as compared to the baseline IOD-KD.

4.4.1 Results on VOC2007 Dataset

We perform 3 different set of experiments with varying number of incremental episodes as follows.

Learning 19+1 Classes. We take 19 classes in alphabetical order from the VOC dataset, and the remaining one as the only new class to be added to the old network (N_{1-19}). Specifically, we first train the the base network N_{1-19} on trainval subset containing any of the 19 classes, and then train the new network N_{20} on the trainval subset containing the only new class. Table 1 shows that our approach outperforms IOD-KD baseline on both the performance measures. While the IOD-KD baseline achieves 58.1% AP on the new class, our approach achieves 59.4% mAP on the new class including an improvement of 0.5% on the old classes. As expected, fine-tuning performs relatively well on the new classes but fails to preserve the accuracy of old classes due to catastrophic forgetting. Our proposed approach on the other hand improves the new class accuracy while preserving the old class accuracy through relation guided knowledge transfer over the selected proposals.

Learning 10+10 Classes. In this experiment, we train the base network on the first 10 classes (alphabetical order) and then use the remaining 10 classes as the new classes. Table 2 shows that our approach outperforms both the Fine-tune and EWC baseline by a significant margin. The IOD-KD baseline is the most competitive. However, we still outperform it by a margin of about 1% in both the measures, showing the utility of object relationships while transferring knowledge from the old to the new network.

Learning 5+5+5+5 Classes. We perform this experiment by using 5 classes at each incremental episode to verify the effectiveness of our approach in multi-episode incremental learning. Table 3 shows the summarized results, with the full results in Table 4. The proposed approach significantly outperforms the IOD-KD baseline by a margin of more than 5% in mAP at the end of the last episode learning. As expected, the performance difference between our approach and the IOD-KD baseline approach increases with the increase in number of episodes. The mAP difference between our proposed method and IOD-KD baseline after the first episode of training is only 0.4% but the difference is about 6% at the end of the episode. IOD-KD fails to preserve the old class accuracy when the number episodes keep increasing. However our approach on the other hand better preserves the old class accuracies by exploiting object relations. For example, we improve by 10% in mAP over the IOD-KD baseline while preserving the accuracy of base 5 classes at the end of the incremental learning (49.9% vs 59%). As seen in Table 4, for classes like *bird* we improve by about 25% in AP over IOD-KD at the last episode of the learning by selecting proposals that overlap with the ground truth *plant* class. We believe this is due to the fact that *bird* and *plant* often co-occur in the same image.

4.4.2 Results on VOC2012 Dataset

We perform two different set of experiments one with single episode (10+10) and another with 3 episodes (5+5+5+5) to compare with different methods.

Learning 10+10 Classes. Table 5 summarizes the results. Similar to the results in VOC07 dataset, the proposed ap-

Method	mAP_{1-5}	mAP_{6-10}	mAP_{11-15}	mAP_{16-20}	mAP	Ω_B
N_{1-5}	57.6	-	-	-	-	-
N_{1-5} + N_{6-10} w/ IOD-KD [37]	60.5	51.4	-	-	55.9	0.85
N_{1-5} + N_{6-10} w/ RKT (Ours)	60.1	53.5	-	-	56.8	0.86
N_{1-10}	64.7	66.6	-	-	65.6	1.00
N_{6-10} + N_{11-15} w/ IOD-KD [37]	56.7	47.6	56.8	-	53.7	0.77
N_{6-10} + N_{11-15} w/ RKT (Ours)	61.1	51.6	57.8	-	56.9	0.82
N_{1-15}	64.9	69.7	73.9	-	69.5	1.00
N_{11-15} + N_{16-20} w/ IOD-KD [37]	49.9	41.7	55.0	41.6	47.0	0.69
N_{11-15} + N_{16-20} w/ RKT (Ours)	59.0	49.5	57.6	45.4	52.9	0.77
N_{1-20}	65.0	70.9	73.3	64.1	68.3	1.00

Table 3. **Learning 5+5+5+5 Classes in VOC07 Dataset.** Results show multiple episodic performance of different methods while adding 5 classes at each episode of the incremental learning. We also report the base class performance of a network trained using the same number of classes without any incremental learning. Our proposed approach performs the best, especially when the number of episodes keep increasing. We outperform IOD-KD by a margin of about 9% in mean mAP on the old 15 classes at the end of incremental learning.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hor	mbi	per	plant	sheep	sofa	train	tv
N_{1-5}	69.1	69.7	48.7	52.3	48.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
+ N_{6-10} w/ IOD-KD [37]	70.6	73.3	56.6	54.1	48.1	52.5	67.9	57	26.2	53.3	-	-	-	-	-	-	-	-	-	-
+ N_{6-10} w/ RKT (Ours)	68.2	70.8	54.5	53	54.0	53.3	67.3	62.9	32.3	51.5	-	-	-	-	-	-	-	-	-	-
+ N_{11-15} w/ IOD-KD	69.7	70.1	44.5	51.9	47.4	50.8	66.9	42.5	30.0	47.9	44.6	51.2	67.2	60.4	60.7	-	-	-	-	-
+ N_{11-15} w/ RKT	68.5	74.4	56.1	53.8	54.7	54.8	68.0	59.0	32.3	43.8	50.1	53.7	61.5	60.0	63.9	-	-	-	-	-
+ N_{16-20} w/ IOD-KD	63.3	70.1	26.8	47.9	41.2	40.6	67.3	42.7	30.2	27.6	44.8	46.7	64.3	63.6	56.0	26.2	33.5	41.9	51.6	54.7
+ N_{16-20} w/ RKT	68.0	74.2	50.2	53.1	49.7	53.1	68.5	50.0	34.1	41.9	52.0	51.0	59.5	60.9	64.8	27.2	46.1	46.4	47.0	60.2
N_{1-20}	75.6	77.6	68.0	55.2	48.6	75.7	79	77.7	44.8	77.3	65.8	75.3	79.9	69.7	75.9	43.6	65.2	66.1	76.4	69.3

Table 4. **Per-class Performance of Learning 5+5+5+5 Classes in VOC07 Dataset.** Results show per-class average precision while adding 5 classes at each episode of the incremental learning. We also report the base class performance of a network trained using the same number of classes without any incremental learning. Our approach on an average, outperforms the baseline on most of the classes.

Method	mAP_{1-10}	mAP_{11-20}	mAP	Ω_B
N_{1-10}	60.3	-	-	-
+ N_{11-20} w/ Fine-tune	5.90	56.7	31.3	0.62
+ N_{11-20} w/ IOD-KD [37]	48.5	52.1	50.3	0.79
+ N_{11-20} w/ RKT (Ours)	57.8	49.4	53.6	0.84
N_{1-20}	63.4	63.4	63.4	1.00

Table 5. **Learning 10+10 Classes in VOC12 Dataset.** Results show the addition of 10 classes, all at once, to a pre-trained object detection network initially trained with 10 classes. Our approach outperforms the baseline method by a margin of 3.3% in mAP.

proach outperforms both Fine-tune and IOD-KD baselines by a significant margin. The Fine-tuning baseline performs very poorly on the old classes, reaching only 5.90% mAP compared to the 57.8% accuracy achieved using our method for the old classes. We improve over IOD-KD by 3.3% in overall mAP by selecting both right proposals and relations while transferring knowledge from the old to the new network. Note that while our performance on the new 10 classes are about 3% lower than IOD-KD, we improve by more than 9% in old class accuracy which once again show the efficacy of our method in reducing the catastrophic interference in incremental learning.

Learning 5+5+5+5 Classes. Results of adding 5 classes at each episode are shown in Table 6 and Table 7. Similar to the results in VOC07 dataset, our approach consistently outperforms the IOD-KD baseline at each episode of the learning. We observe that the performance of both old and new classes are well preserved using our approach that not only focuses on selecting the right knowledge but also on how to effectively transfer them for minimizing the effect of catastrophic forgetting in incremental object detection.

4.4.3 Results on KITTI Dataset

Learning 2+1 Classes. In this experiment we train the base network on the first 2 classes (*Car* and *Cyclist*) and then use the 3rd class (*Pedestrian*) for incremental learning. Table 8 shows that our approach outperforms the IOD-KD baseline on both new and old accuracies. This once again shows the utility of object relationships for transferring knowledge even in datasets with extremely small number of classes.

4.5. Ablation Analysis

To understand the impact of the different components, we analyzed the performance of the proposed approach, by ablating each component on the VOC07 dataset. With all the components working, the mAP while learning 10 classes incrementally is 63.1%. By turning off our proposed proposal selection strategy (i.e., no ground truth bounding boxes are used for selecting proposals), the mAP decreases

Method	mAP_{1-5}	mAP_{6-10}	mAP_{11-15}	mAP_{16-20}	mAP	Ω_B
N_{1-5}	51.1	-	-	-	-	-
N_{1-5} + N_{6-10} w/ IOD-KD [37]	53.7	48.1	-	-	50.9	0.77
N_{1-5} + N_{6-10} w/ RKT (Ours)	54.4	48.2	-	-	51.3	0.78
N_{1-10}	64.7	66.6	-	-	65.6	1.00
N_{6-10} + N_{11-15} w/ IOD-KD [37]	54.2	47.0	52.3	-	51.1	0.80
N_{6-10} + N_{11-15} w/ RKT (Ours)	56.2	46.8	54.8	-	52.6	0.82
N_{1-15}	61.0	62.5	68.4	-	63.9	1.00
N_{11-15} + N_{16-20} w/ IOD-KD [37]	53.3	44.6	51.0	36.6	46.4	0.73
N_{11-15} + N_{16-20} w/ RKT (Ours)	55.5	42.6	52.0	39.3	47.3	0.75
N_{1-20}	61.5	65.3	69.6	57.3	63.4	1.00

Table 6. **Learning 5+5+5+5 Classes in VOC12 Dataset.** Results show multiple episodic performance while adding 5 classes at each episode of incremental learning. Our approach outperforms the baseline at each episode of the incremental learning.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hor	mbi	per	plant	sheep	sofa	train	tv
N_{1-5}	78.3	60.7	37.1	34.9	44.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
+ N_{6-10} w/ IOD-KD [37]	78.6	64.7	49.3	29.6	46.2	60.6	53.7	72.5	24.2	29.2	-	-	-	-	-	-	-	-	-	-
+ N_{6-10} w/ RKT (Ours)	79.6	65.2	47.7	31.5	48.0	55.3	53.8	73.7	26.1	32.0	-	-	-	-	-	-	-	-	-	-
+ N_{11-15} w/ IOD-KD	77.1	67.9	49.8	31.7	44.5	60.1	54.6	70.2	23.7	26.4	30.6	60.6	41.4	63.6	65.4	-	-	-	-	-
+ N_{11-15} w/ RKT	79.4	69.6	51.5	32.9	47.4	55.9	53.6	72.5	21.6	30.4	38.2	62.4	44.1	61.9	67.3	-	-	-	-	-
+ N_{16-20} w/ IOD-KD	75.8	67.4	45.2	34.2	43.9	54.7	52.9	65.3	22.8	27.5	28.2	62	39.8	64.8	60.4	21.0	44.8	28.6	31.3	57.4
+ N_{16-20} w/ RKT	77.3	69.9	47.8	35	47.3	52.9	49.3	62.5	17.4	31.0	33.6	60.9	44.5	66.4	54.5	22.6	44.1	30.7	39.9	59.2
N_{1-20}	79.4	71.8	67.9	43.2	45.6	75.5	65.1	85.3	41.7	57.9	49.7	82.4	68.4	73.2	74.3	37.1	60.5	48.9	73.1	67.0

Table 7. **Per-class Performance of Learning 5+5+5+5 Classes in VOC12 Dataset.** Results show per-class average precision while adding 5 classes at each episode of the incremental learning. Our proposed approach outperforms the baseline on most of the classes.

Method	mAP_{1-2}	mAP_3	mAP	Ω_B
N_{1-2}	42.0	-	-	-
+ N_3 w/ IOD-KD [37]	44.7	33.2	40.9	0.97
+ N_3 w/ RKT (Ours)	45.2	34.6	41.7	0.99
N_{1-3}	44.1	38.1	42.1	1.00

Table 8. **Learning 2+1 Classes in KITTI dataset.** Results show addition of 1 class to base network trained with two classes. Even in this small dataset, our approach outperforms IOD-KD.

Method	mAP_{1-10}	mAP_{11-20}	mAP	Ω_B
N_{1-10}	65.8	-	-	-
+ N_{11-20} w/o Proposal Selection	65.3	54.0	59.7	0.87
+ N_{11-20} w/o Relation Transfer	66.6	57.1	61.9	0.90
+ N_{11-20} w/ RKT (Ours)	67.1	59.2	63.1	0.92
N_{1-20}	67.9	68.7	68.3	1.00

Table 9. **Ablation Analysis with 10+10 classes on VOC2007 dataset**. Results of incrementally adding 10 classes (N_{11-20} to base network (N_{1-10}) trained with the first 10 classes under different settings. Our approach works the best while both proposal selection and relation transfer are working.

to 59.7%. This highlights the importance of selecting the right object proposals, i.e information for transfer for incremental learning. Similarly, by turning off the relation transfer (i.e., not using relation loss over selected proposals), the mAP becomes 61.9%. Overall, adding relation transfer loss over the selected object proposals further improves the result on both the old and new classes.

5. Conclusion

In this paper, we propose a relation guided knowledge transfer approach that use proposal relationships for incremental learning of object detectors without accessing old classes data. Specifically, we first propose a proposal selection mechanism that utilizes ground truth as priors for selecting what knowledge to transfer and then introduce a relation guided transfer loss to preserve the relations of selected proposals between the base network and the new network trained on new classes. Extensive experiments show that our approach outperforms the baselines highlighting the importance of relationships in object detection. While we explored incremental learning in the context of images, we believe that our approach can be extended to videos where natural relations between objects and actions exist. We leave this as an interesting direction for future work.

Acknowledgements. This work is in part supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00341. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, 2018.

[2] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *ECCV*, 2018.

[3] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, 2018.

[4] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *NIPS*, 2017.

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017.

[6] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *CVPR*, 2019.

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[8] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2009.

[9] Ross Girshick. Fast r-cnn. In *CVPR*, 2015.

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[13] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Lifelong learning via progressive distillation and retrospection. In *ECCV*, 2018.

[14] Nitin Kamra, Umang Gupta, and Yan Liu. Deep generative dual memory network for continual learning. *arXiv preprint arXiv:1710.10368*, 2017.

[15] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *AAAI*, 2018.

[16] Hyo-Eun Kim, Seungwook Kim, and Jaehwan Lee. Keep and learn: Continual learning by constraining the latent space for knowledge preservation in neural networks. In *MICCAI*, 2018.

[17] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 2017.

[18] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018.

[19] Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Incremental learning with unlabeled data in the wild. *arXiv preprint arXiv:1903.12648*, 2019.

[20] Zhizhong Li and Derek Hoiem. Learning without forgetting. *PAMI*, 2017.

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[24] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *CVPR*, 2019.

[25] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *NIPS*, 2017.

[26] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[27] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019.

[28] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019.

[29] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *ICCV*, 2019.

[30] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017.

[31] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[33] Xiaofeng Ren and Deva Ramanan. Histograms of sparse codes for object detection. In *CVPR*, 2013.

[34] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[35] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[36] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *NIPS*, 2017.

[37] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In *ICCV*, 2017.

[38] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, 2019.

[39] Jasper Uijlings, Stefan Popov, and Vittorio Ferrari. Revisiting knowledge transfer for training object class detectors. In *CVPR*, 2018.

[40] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, 2019.

[41] Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. In *ICCV*, 2013.

[42] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 2017.

[43] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[44] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. *arXiv preprint arXiv:1903.07864*, 2019.

[45] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[46] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.