Few-Shot Visual Relationship Co-Localization
- Supplementary Material-

Vaibhav Mishra*, Mayank Maheshwari*, Revant Teotia*, Anand Mishra
Indian Institute of Technology Jodhpur
{mishra.4,maheshwari.2,trevant,mishra}@iitj.ac.in
∗: equally contributed

Contents

List of Figures

List of Tables

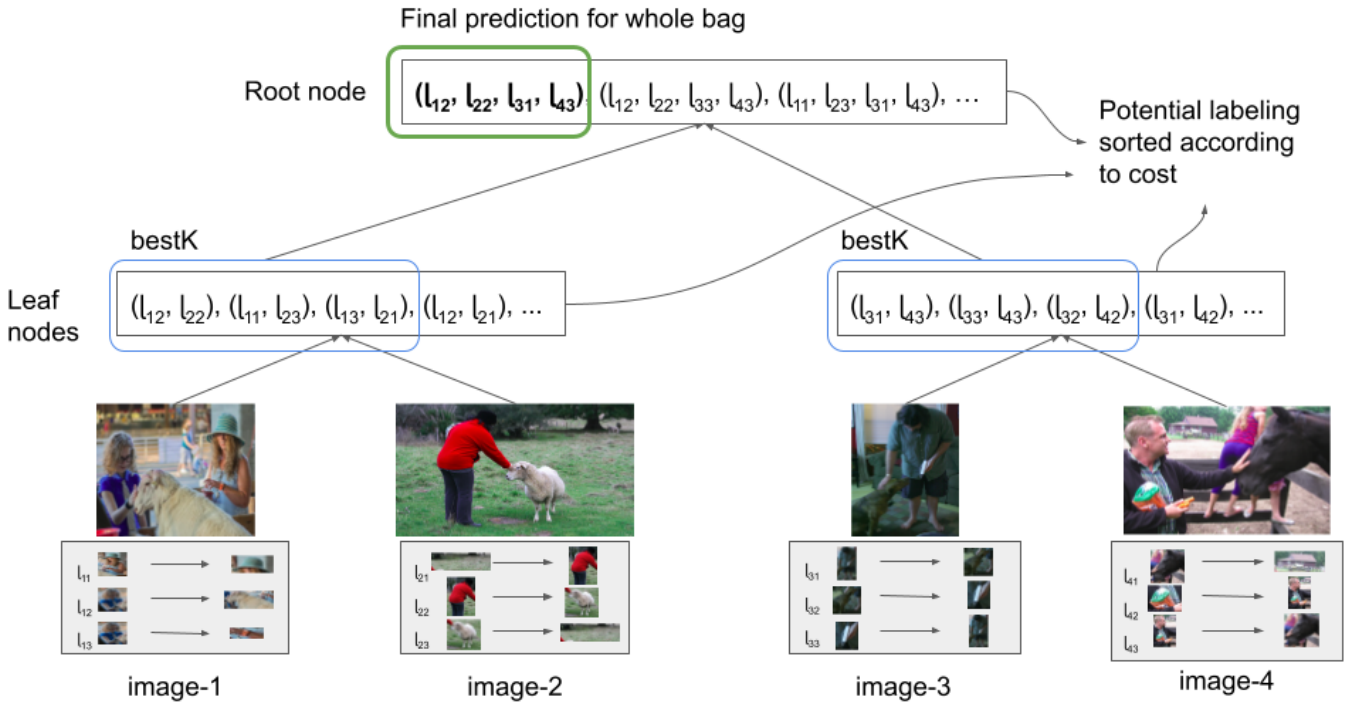


Figure 1. Inference algorithm example for a bag of 4 images. The height of inference binary tree is 1 with two leaf nodes and one root node. For leaf nodes, all possible label pairs of their corresponding images are created and sorted according to their pairwise cost. For the root node, all combinations of the $K = 3$ least cost labeling of its children nodes are created and the one combination with the least cost among them is the final prediction for the whole bag. Blue color boxes represent best K labeling for leaf nodes while green color box for root node represents the final prediction. **[Best viewed in color].**

1. Detailed Inference Algorithm

In this section, we explain inference algorithm presented in Section-3.3 of our main paper in detail. We adopt the greedy approximation algorithm proposed by ref. [26] in the main paper for the problem of visual relationship co-localization (VRC). For a given bag of b images $V = \{I_u\}_{u=1}^{b}$ and its optimal labeling $O = \{l_{ut}\}_{u=1}^{b}$, let $V(p,q) = \{I_u\}_{u=p}^{q}$ be a subset of the bag images, then $O(p,q) = \{l_{ut}\}_{u=p}^{q}$ would also be an optimal labeling for the subset $V(p,q)$. Since the optimal label selection for a large bag of images would also be an optimal solution for a smaller subset of images from the bag, we divide the large bag into smaller bags, find optimal labeling for them, and then combine them greedily to get the labeling for the complete bag.

Without loss of generality we can assume that the bag size $b = 2^z, z \in \mathbb{N}$. We divide the bag into two disjoint equal bags recursively until we get bags of size 2. Since we consider unary cost of selecting a label for an image to be uniform and only pairwise costs contribute to the total labeling cost, the smallest subproblem in VRC is of bag size 2. Thus the whole problem can be represented as a full binary tree of height $z-1$, where each node represents a subproblem. Let $\mathcal{N}_i^h$ be the i-th binary tree node at height h, then each leaf node $\mathcal{N}_i^0$ at height 0 represents a bag size = 2 subproblem, each intermediate node $\mathcal{N}_i^h$ at height h represents disjoint subproblem of the same bag size = $2^{(h+1)}$, and the root node $\mathcal{N}_1^{z-1}$ at height $z-1$ represents the whole problem.

We begin computation at leaf nodes, finding K least cost labeling (let us call it $bestK$) for them. We then move up the tree, and for each tree node $\mathcal{N}_i^h$ at height h, we combine the labeling $bestK_{2i-1}^{h-1}$ and $bestK_{2i}^{h-1}$ of its two children $\mathcal{N}_{2i-1}^{h-1}$ and $\mathcal{N}_{2i}^{h-1}$ respectively. We consider all possible combinations of $bestK_{2i-1}^{h-1}$ and $bestK_{2i}^{h-1}$ and keep the K least cost labeling among them as $bestK_i^h$. Finally, after getting $bestK_1^{z-1}$ for the root tree node $\mathcal{N}_1^{z-1}$, we take the one with the least cost among them as the final labeling for the whole bag. Please refer to figure 1 for an illustrated example of inference.

Why we keep only bestK labeling for each partial solution during inference?: As explained in section 3.3 of the main paper, the cardinality of the label set of an image is $p_u \times (p_u - 1)$ where p_u is number of object proposals in image-u. For

images in our experiments, p_u is usually between 30-50 making the cardinality of the label set around 1600. Further, during inference as we go up the tree combining the labeling of children nodes, the number of potential labeling for the parent node grows exponentially if we consider all possible labeling of the children. Therefore, it becomes necessary to greedily prune the possible labeling combinations by keeping the $bestK$ otherwise it would take $O\big((\prod_{u=1}^{b} p_u^2)\big)$ time to get the optimal labeling for the whole bag.

2. More Results

In the following section, we show an extensive analysis and results of our proposed approach in both qualitative and quantitative ways.

2.1. Qualitative

2.1.1 No supervision

This is the original problem setting of our paper where we only give the bag of images as input and no other kind of supervision. We show our results for nine bags of size four in Figure 2, Figure 3 and Figure 4. Note that in all the images green bounding box refers to predicted subject and yellow bounding box refers to predicted object.

2.1.2 Subject fixed in all images of bag

In this problem setting, as explained in manuscript also, we relax a supervision constraint to evaluate the performance of our proposed approach in these cases also. In this setting we input a bag of images along with the subject bounding boxes for all the images of the bag. The output of the proposed approach is shown in Figure 5 and Figure 6 with bag size of four and eight respectively.

2.1.3 Subject-Object pair provided in one image

In this problem setting, we relax the condition of no-supervision by providing the ground truth subject-object bounding box of only one image in the bag. Therefore, effective input would be a bag of images with first image containing both bounding boxes for subject and object. The results for this is shown in Figure 7, Figure 8, Figure 9. It should be noted that we provide bounding box pair annotation for only the first image of the row in each case.

2.1.4 Failure cases

We also show in Figure 10 a few cases where our approach fails or does not select the right subject or object which is connected by the hidden predicate (written below each image for reference).

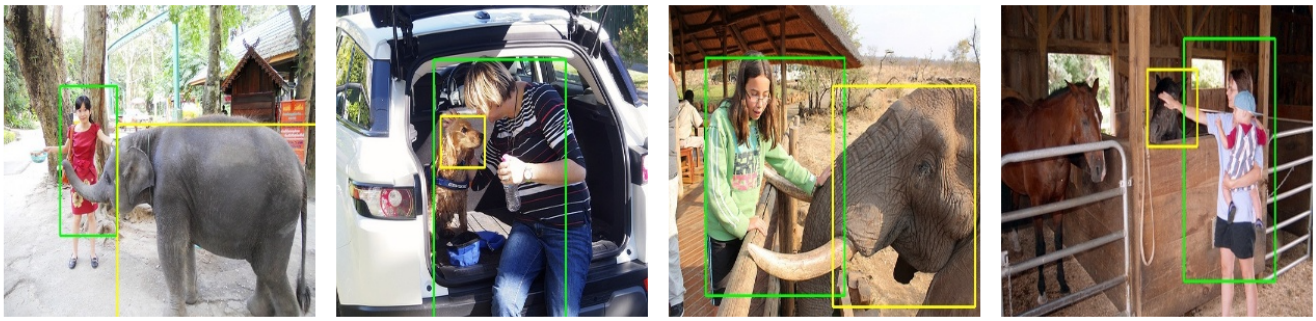
2.2. Quantitative

2.2.1 Mean IoU of Visual Relationship Co-Localization for VrR-VG

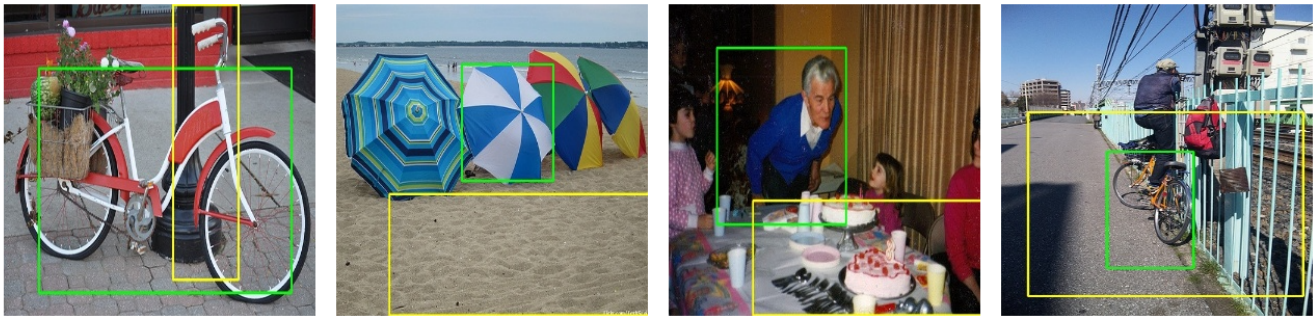
To reinforce the robustness of our proposed algorithm against baseline approaches, we also calculated Mean-Intersection over Union (mean-IoU) between ground truth visual subject-object bounding boxes and our predicated bounding boxes of all the images in the test set of Vrr-VG dataset, and report the results in Table 1. We observe that even on this performance measure, our approach outperforms each of the baseline by a significant margin.

2.2.2 Class-wise performance for 17 test predicates of VrR-VG

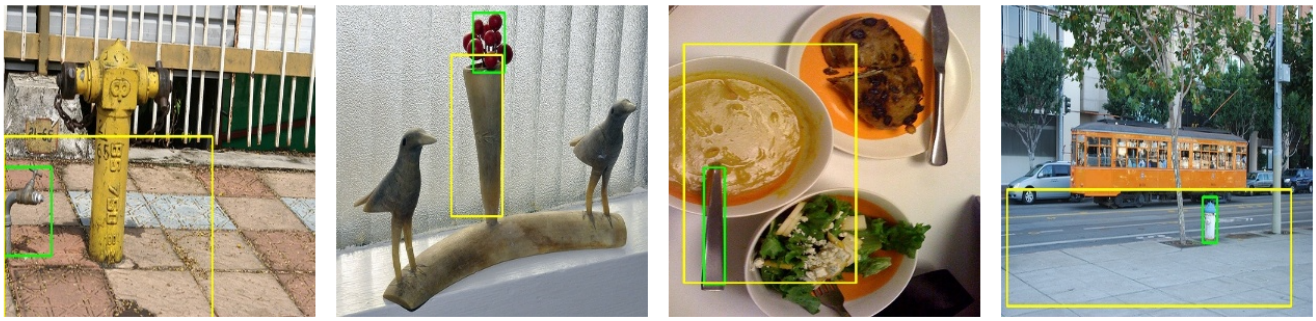
For further extensive analysis, we also calculate both VR-CorLoc and Bag-CorLoc on all the test predicates of the VrR-VG dataset individually. All of these 17 test predicates were only used for testing purpose i.e. they were never seen before by our approach. With this analysis we get to know that on which predicates our approach is performing better than other so that we can improve upon all the low performing classes. Results of this analysis can be easily inferred from table 2. We can see that 'petting', 'sniffing', 'following' are some of the top performing predicates.



Petting

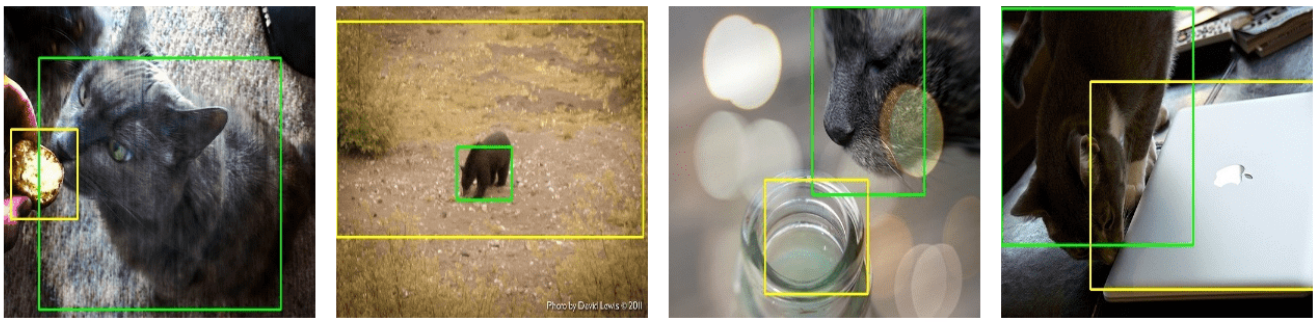


Leaning on

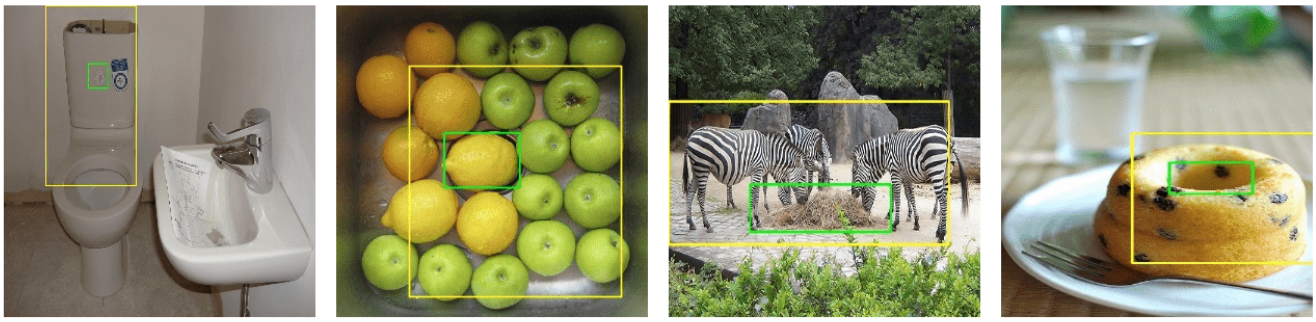


Sticking out of

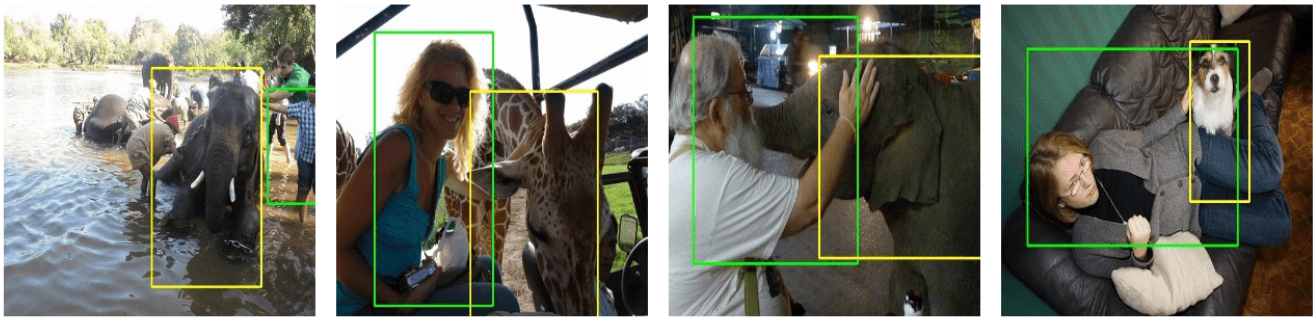
Figure 2. Figure has output of our proposed approach on three **testing predicates (unseen before)**. Each row corresponds to a bag of images with **bag size as four** and each of the latent predicate of the bag is written below each row (equivalently bag). Note that green bounding box is for predicted subject and yellow is for predicted object. **[Best viewed in color].**



Sniffing

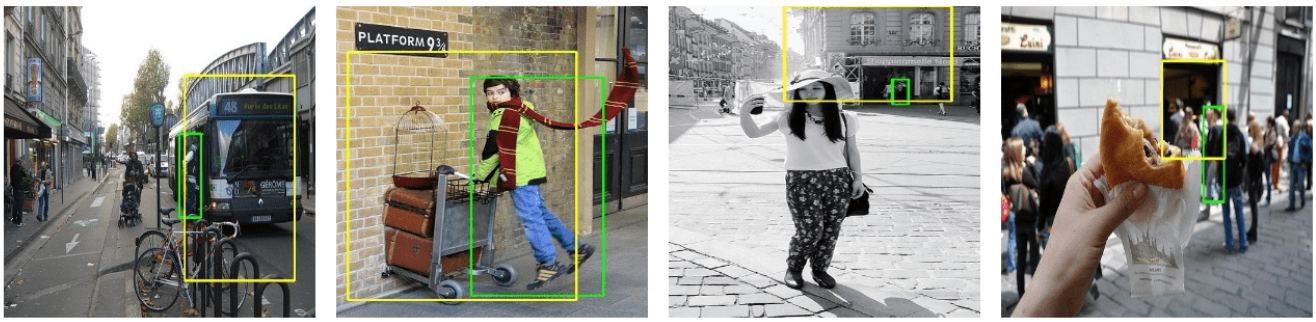


In Center of



Petting

Figure 3. Continuing from previous figure we show output of our proposed approach on three **testing predicates (unseen before)**. Each row corresponds to a bag of images with **bag size as four** and each of the latent predicate of the bag is written below each row (equivalently bag). Note that green bounding box is for predicted subject and yellow is for predicted object. **[Best viewed in color].**

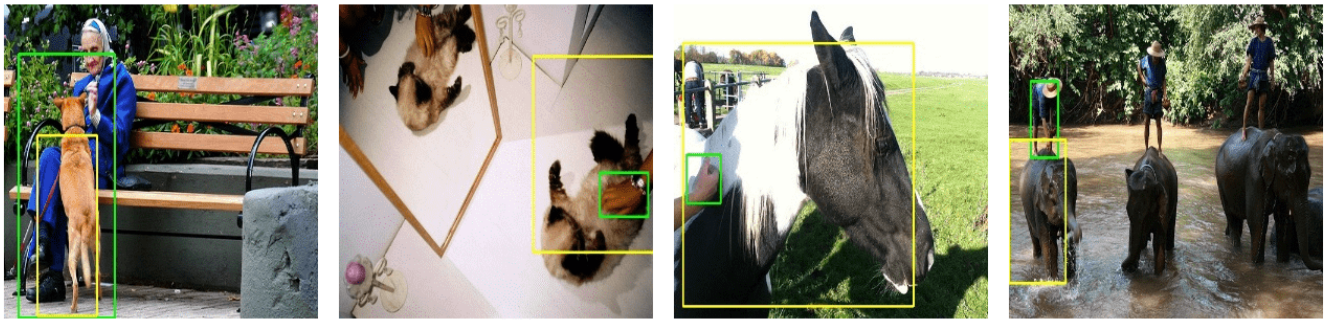


Entering

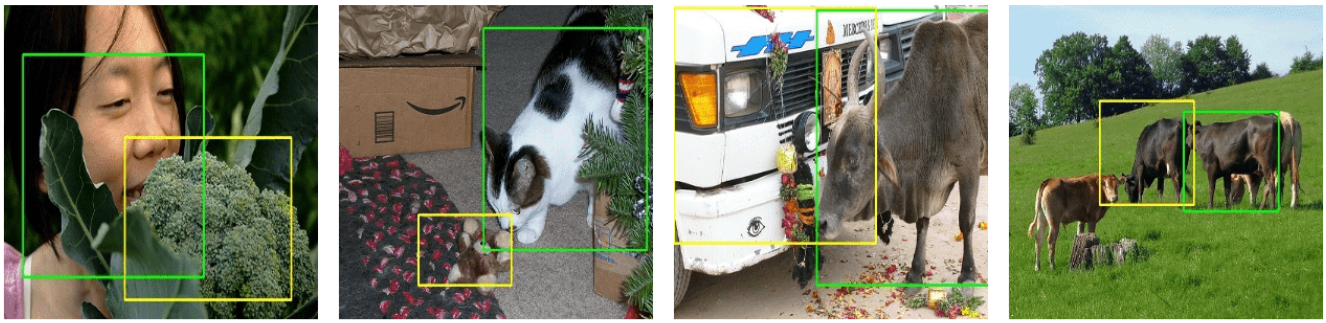
Sniffing

Sticking out of

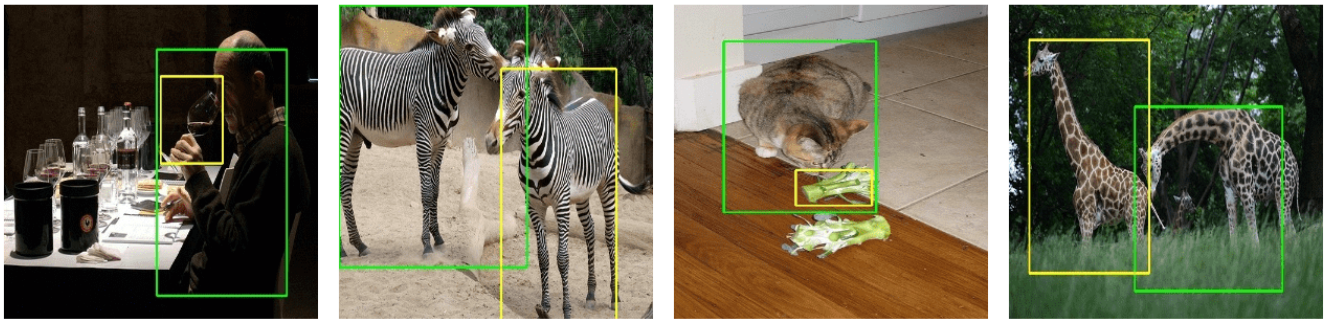
Figure 4. Figure has output of our proposed approach on various **testing predicates (unseen before) in No Supervision** setting. Each row corresponds to a bag of images with **bag size as four** and each of the latent predicate of the bag is written below each row (equivalently bag) so as we can interpret the results. Also, Note that green bounding box is for predicted subject and yellow is for predicted object. **[Best viewed in color].**



Petting



Sniffing



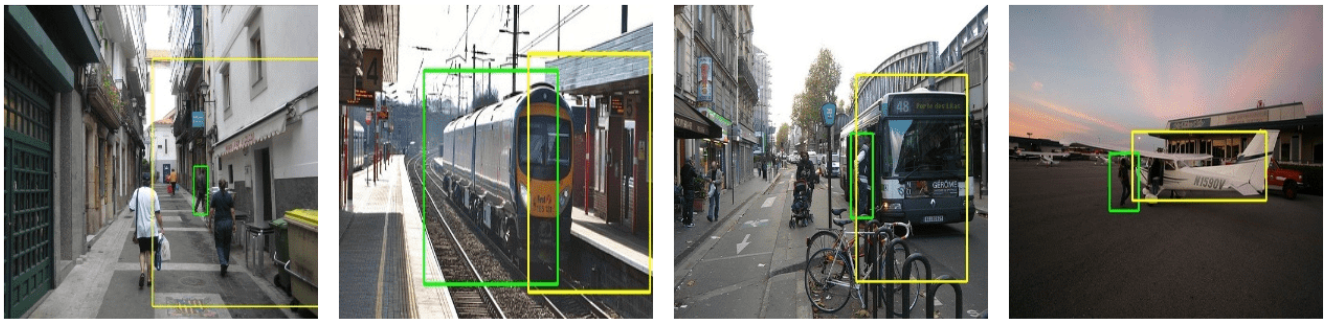
Sniffing

Figure 5. Figure has output of our proposed approach on various **testing predicates (unseen before) and the Subject boxes given for all images** setting. Each row corresponds to a bag of images with **bag size as four** and each of the latent predicate of the bag is written below each row (equivalently bag) so as we can interpret the results. Also, Note that green bounding box is for predicted subject and yellow is for predicted object. **[Best viewed in color].**

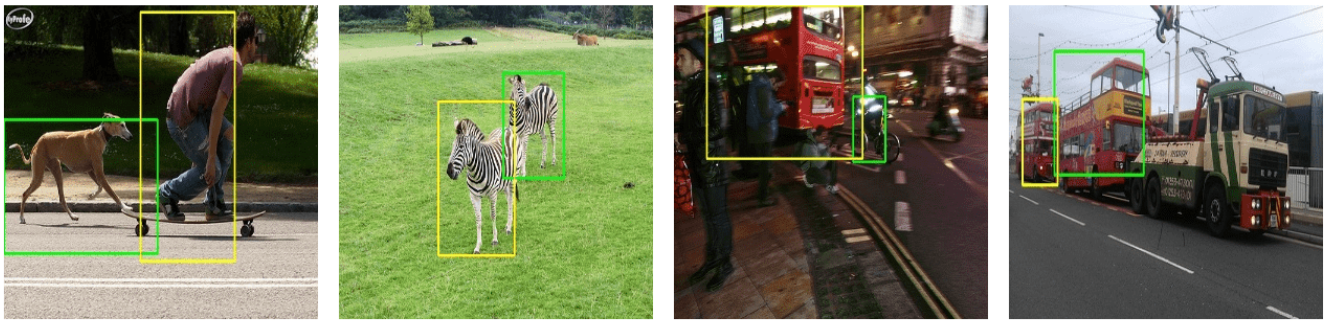
Figure 6. Figure has output of our proposed approach on various **testing predicates(unseen before) and the subject bounding boxes are provided for all images** setting. Each row corresponds to a bag of images with **bag size as eight** and the latent predicate of the bag is written below so as we can interpret the results. Also, Note that green bounding box is for predicted subject and yellow is for predicted object. **[Best viewed in color].**

Method	Bag Size	IoU
Concat + Cosine Similarity	2	0.5744
	4	0.5808
	8	0.5778
VTransE + Cosine Similarity	2	0.5920
	4	0.6175
	8	0.6097
Concat + Relation Net	2	0.6215
	4	0.6112
	8	0.6275
Our Approach	2	**0.6463**
	4	**0.6656**
	8	**0.6480**

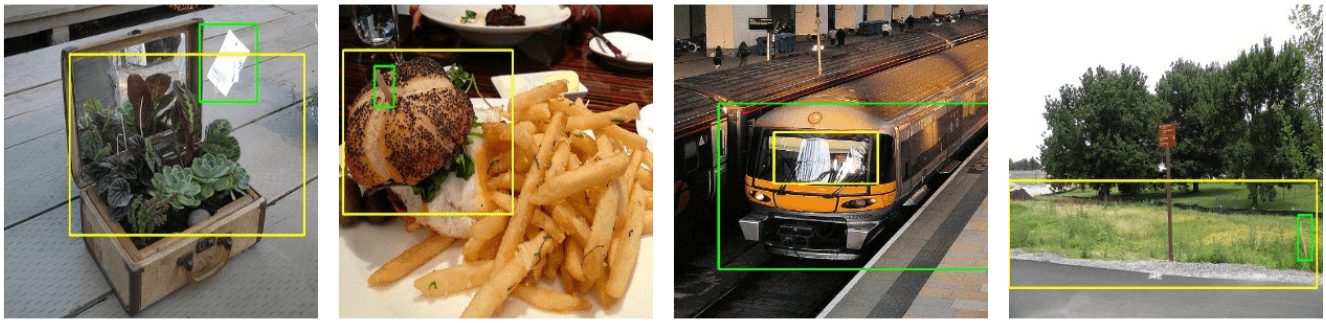
Table 1. **Mean IoU between ground truth visual subject-object bounding boxes and our predicated bounding boxes of all the images in the test set of Vrr-VG dataset. We observe superior performance of the proposed approach under this performance measure as well.**



Entering

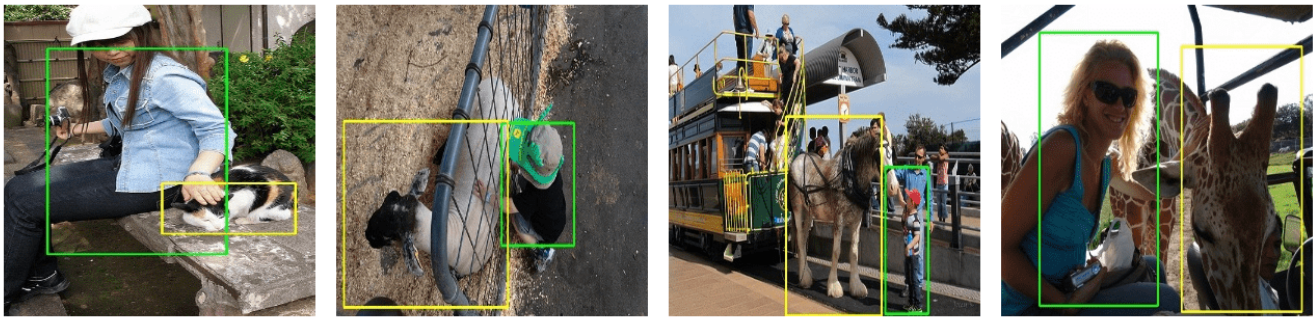


Following

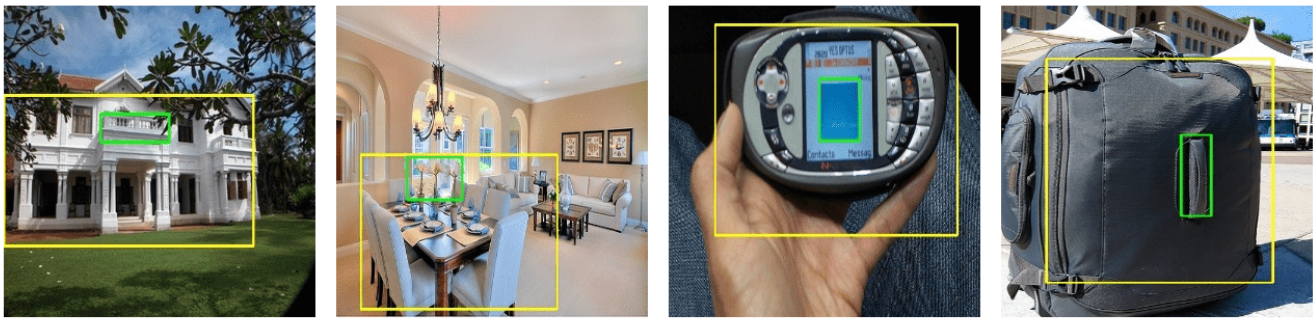


Sticking out of

Figure 7. Figure has output of our proposed approach on various **testing predicates(unseen before) and the Subject-Object boxes given for one image** setting. Each row corresponds to a bag of images with **bag size as four** and each of the latent predicate of the bag is written below each row (equivalently bag) so as we can interpret the results. Also, Note that green bounding box is for predicted subject and yellow is for predicted object.**[Best viewed in color].**



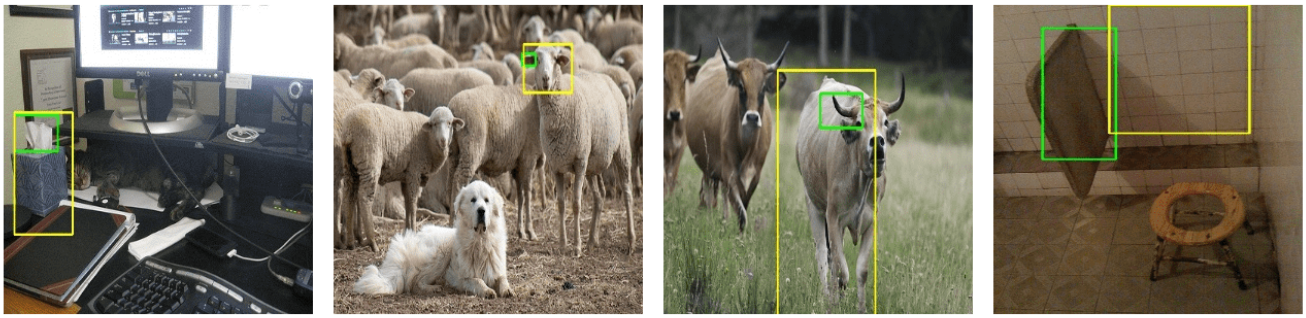
Petting



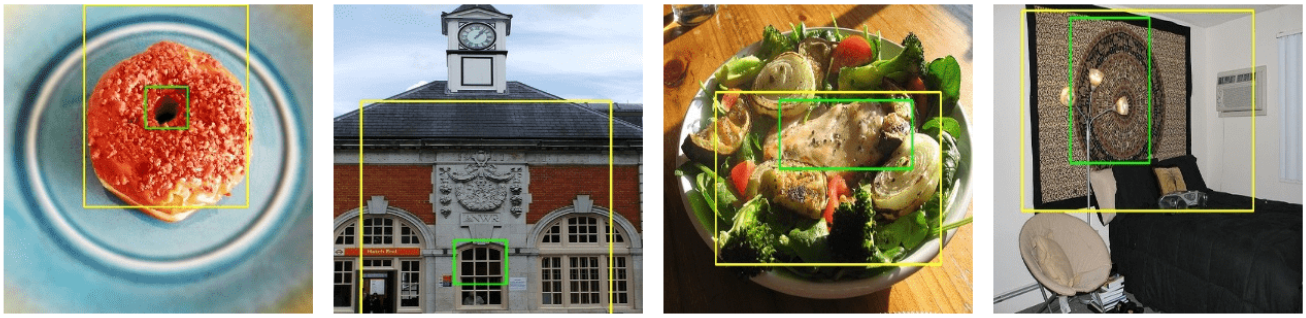
In Center of

Entering

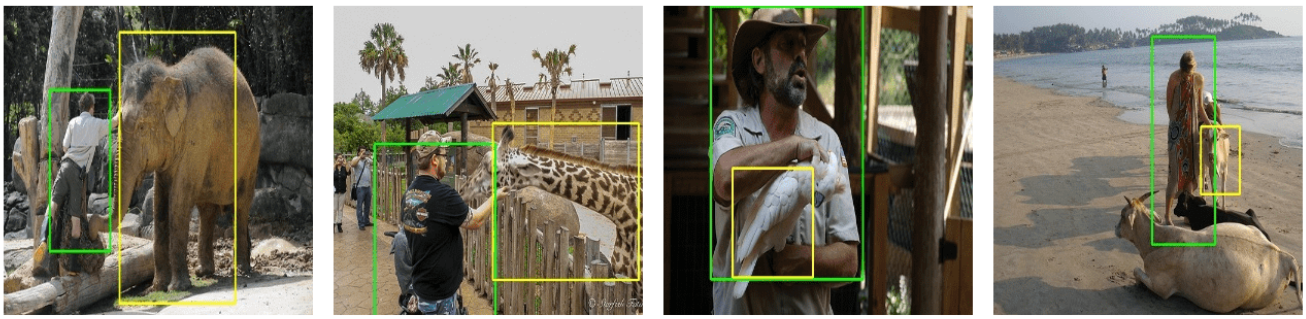
Figure 8. Figure has output of our proposed approach on various **testing predicates(unseen before) and the Subject-Object boxes given for one image** setting. Each row corresponds to a bag of images with **bag size as four** and each of the latent predicate of the bag is written below each row(equivalently bag) so as we can interpret the results. Also, Note that green bounding box is for predicted subject and yellow is for predicted object. **[Best viewed in color].**



Sticking out of



In Center of



Petting

Figure 9. Figure has output of our proposed approach on various **testing predicates(unseen before) and the Subject-Object boxes given for one image** setting. Each row corresponds to a bag of images with **bag size as four** and each of the latent predicate of the bag is written below each row(equivalently bag) so as we can interpret the results. Also, Note that green bounding box is for predicted subject and yellow is for predicted object.**[Best viewed in color].**

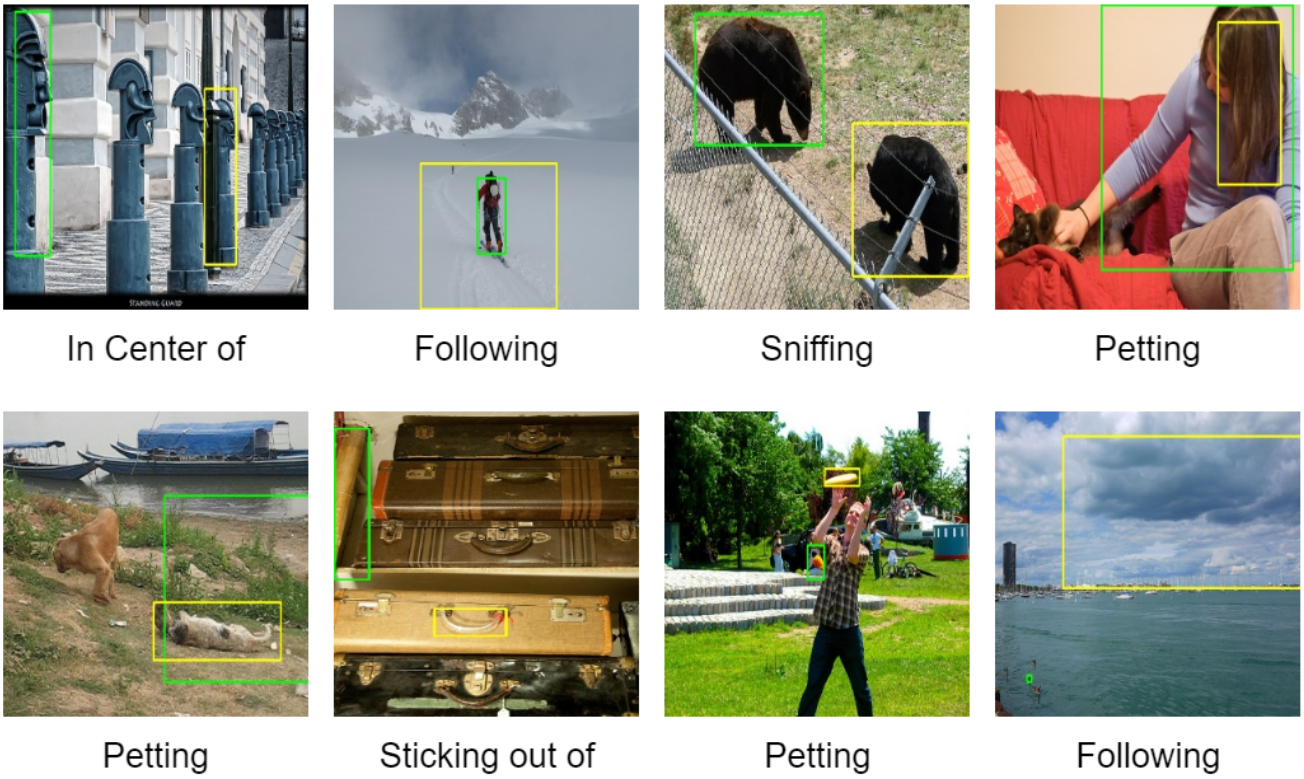


Figure 10. **Failure Cases.** We show some of the output of our proposed approach that varied greatly from the ground truth which in turn led to failure on these images. We can see the hidden predicate written below each of the image for the reference. We observe that sometimes our model predicts a similar visual relationship which is not the ground truth common relationship but semantically very close to it. In some of the other cases our approach seems to localizes one of the subject or object accurately but fails to localize the other counterpart thus failing on whole. Also, it should be noted that green bounding boxes represent predicted subject and yellow represent predicted object. **[Best viewed in color].**

Class(predicate)	Bag-CorLoc (in %)	VR-CorLoc (in %)
biting	38.70	77.9
petting	80.85	94.9
sniffing	66.00	90.25
pointing	1.33	42.66
placed on	24.70	68.23
stacked on	12.50	62.15
balancing on	61.22	88.26
drawn on	9.58	56.84
sewn on	10.20	61.73
sticking out of	13.79	62.93
at bottom of	12.35	59.55
following	66.32	91.07
entering	16.32	63.01
leaning on	34.93	78.90
in corner of	37.00	75.25
surrounded by	24.63	67.39
in center of	42.85	79.84

Table 2. **Class-wise performance of our approach for Visual Relationship Co-Localization on VrR-VG dataset.**