**** Our method:** The proposed co-attention augments the feature maps (for RPN) of a target image I with non-local information w.r.t. the query patch p. The step leads to non-local object proposals. To determine the similarity between the query patch and each proposal, we design a co-excitation mechanism to simultaneously re-weight and obtain the feature maps $\tilde{F}(I)$ and $\tilde{F}(p)$. The co-excited feature representations are coupled with a margin-based ranking loss to uncover instances similar to the query, no matter its category is seen or unseen. We emphasize that our method does not require any model fine-tuning for carrying out the task of one-shot object detection over the unseen object categories.

**** Reviewer #1**

1.1 *"The paper extends the previous work [24]... Other components of the detector are standard. I cannot name the..."*
While the design to generate non-local object proposals is new and its usefulness is illustrated in Figure 3, the other main idea of our approach is to link the optimization of a margin-based ranking loss with co-excited features. As demonstrated in the experiments (e.g., Figures 5 & 6), the formulation enables our detection method to uncover instances of either a seen or an unseen class that resemble the query patch in different feature aspects such as shape and texture.

1.2 *"The paper addresses a special case of few-shot object detection, subject to the following restrictions..."*
One-shot object detection is the most challenging one among all k-shot ($k \geq 1$) detection scenarios. The assumption that the target image includes at least one object instance w.r.t. the class label of the query is assumed mainly for training. In testing, although we have adopted the same tactic but it is for the sake of comparison with other techniques. After all, in inference the detection results are subject to a unified score threshold (which is 0.5 in our implementation). The one-shot query in training is indeed to mimic the situation of one-shot (unseen class) object detection in inference. Notice that in our formulation those single examples from unseen classes are provided only in inference and unlike [20, 21, 23], our method does not require any model fine-tuning for carrying out the one-shot (unseen) object detection.

1.3 *"Here its textual description contradicts the diagram... reweighted feature vector $\widetilde{F}(I)$ is the input to RPN, but..."*
The reviewer may have missed the up and down arrows from the Channel Attention box in Figure 1 which conform to $\mathrm{SCE}(F(p), F(I)) = \mathbf{w}$. On lines 130–134, we explain that $F(I)$ is the input to RPN, as in Figure 1. The SCE reweighted features $\tilde{F}(I)$ are instead coupled with the margin-based ranking loss. Also please see our method above.

1.4 *"I believe [26] deserves more credit for the squeeze-and-excitation idea..."*
The SE block [26] is now a popular technique to re-weight the feature maps. In our writing we have explicitly stated that our implementation of SCE follows [26]. However, unlike the SE block that works on feature maps from a single source, the proposed co-excitation of SCE involves two streams of feature maps from two different input sources.

1.5 We thank the reviewer for pointing out the mistake in misplacing "bus" and "person" in Table 1.

**** Reviewer #4** (Due to the limited space allowed, we have chosen to response to those most critical questions or concerns.)

4.1 *"RPN using co-attention is superior to a simple class-agnostic RPN that does not depend on the query image..."*
We have provided qualitative results in Figure 3 and ablation evaluation in Table 3. As noticed by the reviewer, it is difficult to quantitatively analyze the effect of RPN with or without performing the non-local operation w.r.t the query patch in that the co-attention feature maps are used not only in RPN but also in the later stages of the network.

4.2 *"The ranking loss in equations 5 and 6 is mostly logical except..."*
Since the first two terms in the RHS of (5) already enforce that foreground proposals would have larger scores than those of the background ones, we emphasize only the score difference between each proposal pair depending on whether they have the same class label. In (5) and (6), we have used the same m^+ and m^- simply for reducing the number of margin parameters from 4 to 2. We suspect that the all-pair ranking loss is not new but need to strive to find a relevant citation.

4.3 *"The form of the squeeze function SCE() is not stated explicitly..."*
The squeeze step of SCE is performed only with $F(p)$. The co-excitation $\mathbf{w}$ involves two streams of feature maps (see Figure 1) and re-weights $F(I)$ and $F(p)$ into $\tilde{F}(I)$ and $\tilde{F}(p)$. Our method learns to find an appropriate co-excitation such that the margin-based ranking loss could prefer foreground proposals. Also see our response in 1.4.

4.4 *"You should cite the journal version of [26]... In Figure 1, tilde{F}(I) is not connected to anything?..."*
Thank you and we will cite the journal version of [26]. We are sorry about the confusion in Figure 1. We have duplicated $\tilde{F}(I)$ and positioned it after the RPN to indicate that the non-local proposals are represented with features from $\tilde{F}(I)$.

**** Reviewer #5**

5.1 *"concern about the evaluation protocol: at test time the query image fed is from the classes present in target image"*
We agree with the reviewer that the current evaluation protocol can be improved. Still, it is adopted by most of the related work about one-shot object detection, e.g., [22, 24]. As suggested by the reviewer, a more insightful evaluation protocol, e.g., training from animal meta-categories and one-shot testing on vehicle meta-categories seems to pose a very challenging problem of one-shot object detection, but we would attempt to carry out the suggested explorations and include such experimental results for thoroughly testing the effectiveness of our proposed method.