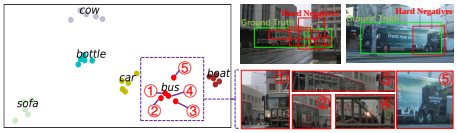

Figure 1: Distribution of embedded negative representatives using t-SNE; The selected hard negative proposals from support images (3-shot, i.e. 3 *buses*) of *bus*. [Zoom In]

Shot	1	2	3	5	10
Novel Set 1	26.1	32.9	34.4	38.6	41.3
Novel Set 2	17.2	22.1	23.4	28.3	35.8
Novel Set 3	27.5	31.1	31.5	34.4	37.2

Table 1: Performance of RepMet on PASCAL VOC novel sets.

We thank all reviewers for their constructive and valuable comments.

(R1) Baseline: training a simple object detector on base classes and fine-tuneing its head on novel classes.
A1: Indeed, the proposed baseline has already been implemented in RepMet[17] (see its Page 6 and Table 3 for "baseline-FT"). This baseline on ImageNet-LOC performs much inferior than RepMet/NP-RepMet. We will add this.

(R1) The backbone and pre-training regime of the network should be controlled for comparison.
A2: We agree ! We mentioned this in Sec. 4.2: we use ResNet-101 as the backbone (L213); its weights initialization follows [17] for ImageNet-LOC and [16] for PASCAL VOC (L215); weights of other modules (e.g. FPN, RPN) are randomly initialized (L216). That is, we pre-train the backbone on COCO to fairly compare with RepMet on ImageNet-LOC; pre-train it on ImageNet to align with [16, 36,38] on PASCAL VOC. We are also very careful to follow train/test details (class splits (L211-212), support sets (L224-226), inference scheme (L311-319)) in [17] and [16].

(R2) [16,36,38] do not rely on representatives and hence do not suffer from the problem addressed in this work.
A3: The key observation of this work is that hard negative information within support images is not carefully exploited in previous works [16,17,36,38]. They simply extracts positive proposals w.r.t. ground truth from support images while negatives proposals containing e.g. partial objects or ambiguous surroundings are not considered (L37). These negatives however are important false positives in object detection (L40-43). Although we build our work on RepMet, our idea of restoring negative information should be essential and beneficial to many few-shot detection works. Judging from the results, NP-RepMet drastically outperforms other few-shot methods [16, 36, 38] on PASCAL VOC (e.g. up to 18%).

(R2) It is less intuitive that negative representatives can be used in a query image with different background.
A4: First, we emphasize that hard negative proposals mined in this work are mainly focused on proposals containing partial, occluded object or entire object with massive surroundings. To realize this, we propose to specifically choose them via IoU thresh (0.2, 0.3) and Cluster-Min. Fig. 1 illustrates the selected hard negatives from the support set of *bus*. We also use t-SNE to draw the embedded negative representatives from the support sets of multiple classes: different classes are clearly distinguished from each other; given a query, its hard negative proposals (e.g. partial object) for certain class can be easily filtered out by comparing to those negative representatives from the support set. We provide ablation study in Sec 4.3 and Table 2 to show that these negative representatives contribute substantially to NP-RepMet.

(R2) Discussion on whether the pre-trained data include samples of the novel classes.
A5: Our pre-training follows previous works (see R1-A2). Class overlap between the pre-trained data and novel data may exist, noting that ImageNet contains 1000 classes. Yet, we argue 1) the class distribution on pretrained data is very different to that on novel data; 2) detection-related modules in NP-RepMet are randomly initialized; 3) as long as base class data are enough, we do not think using a pretrained model would cast a strong impact on the final performance.

(R2) More refs. A6: Thanks! We will carefully discuss them! Particularly, the NeurIPS paper indeed has very different train/test settings (class splits, support sets, queries, etc.) to ours and [16,38,36], so the results are not comparable.

(R3) The proposed method is not restricted to few-shot regime, it can be used to train a standard detector
A7: Yes! This can be validated from results in Table 1, 4 in the paper and Table 1 in the supp, where we train the detector on the train set of base classes and evaluate it on the test set of base classes to show it maintains a good accuracy; some comparison to standard detectors can also be found in Table 3 in [38]; notice our NP-RepMet is inferred for all the classes together on PASCAL VOC (L315-319), like a standard detector. On the other hand, if R3 meant fine-tuning NP-RepMet with novel classes instead of meta-testing as in the paper, yes, we can further fine-tune it and our results can be improved to 71.3, 78.6, and 81.7 for 1, 5, and 10-shot on ImageNet-LOC unseen (novel) classes.

(R3) Performance of RepMet on PASCAL VOC.
A8: RepMet did not report results on PASCAL VOC novel sets. We can reproduce it under the same setting with ours, results are in Table 1: RepMet performs clearly inferior to our NP-RepMet (Table 3 in the paper).

(R4) The authors did not report the results of [16] for 3-shot base classes.
A9: It is because [16] did not report it in their paper. One might find it being reported in [38] as 64.8 which is lower than our 66.6; but we also found that [38] reports the 10-shot result of [16] as 63.6 instead of 69.7. Overall, the base class results among NP-RepMet and [16,38] are competitive to each other and all maintain a good accuracy.

(R4) IoU hyper-parameter. A10: Taking IoU>0.7 for positive samples is a common practice for many detection works, e.g. Faster R-CNN and RepMet; we simply follow this. We experimented with IoU>0.5, the result is 68.3 (v.s. 68.5 for IoU>0.7, very close) on the 1-shot case of ImageNet-LOC. As for negative sample selection, we perform ablation study for different IoU in Table 2 (right bottom), this hyper-parameter is quite robust within certain range.

(R4) No comparison to other hard negative mining methods, though they have not applied to few-shot detection.
A11: One representative hard negative mining strategy OHEM [30] is indeed used in RepMet (L101) where class representatives from different classes are considered as negatives to each other. We actually adopt this in our loss function (Eq 1 and 2). We bootstrap the classifier with negative information both within and across images and classes. More importantly, we show that the idea of mining hard negatives within the same image of positives is essential to few-shot detection. We will add more discussion about hard negative mining in literature in the revised version.