We thank all reviewers for their constructive and valuable comments.

(R1) Typos and codes. We will fix typos in the revision and make the code publicly available.

(R1&R4) Evidence for strength of different object representations. We test the detailed metrics for 4 typical models of different representations under similar overall AP (through different backbones), i.e., Faster R-CNN (proposal, AP = 41.0, AP$_{50}$ = **61.3**, AP$_{90}$ = 16.1, AP$_S$ = 24.0, AP$_L$ = 53.5), RetinaNet (anchor, AP = 40.8, AP$_{50}$ = 60.5, AP$_{90}$ = 14.6, AP$_S$ = 22.9, AP$_L$ = 54.6), FCOS (center, AP = 40.9, AP$_{50}$ = 60.3, AP$_{90}$ = 14.3, AP$_S$ = **24.7**, AP$_L$ = 52.3) and CornerNet (corner, AP = 40.4, AP$_{50}$ = 56.2, AP$_{90}$ = **23.4**, AP$_S$ = 20.2, AP$_L$ = **56.3**). We can see that the bounding box representation (Faster R-CNN) is more friendly for classification (highest AP$_{50}$ = 61.3); center representation (FCOS) is more friendly to small objects (highest AP$_S$ = 24.7), and corner representation (CornerNet) is more accurate for larger objects and finer localization (highest AP$_{90}$ = 23.4 and highest AP$_L$ = 56.3).

(R1) Explanation of "works in-place", "part corners" and "evolution flow". The proposed module can be directly plugged into the existing detectors mentioned in the paper. We agree "corners" is a better description. We will also use "representation flow" instead of "evolution flow" as suggested. These terms will be modified in the revised version.

(R1&R4) Intuitiveness of Section 3.1. We will rewrite it to be more intuitive and more friendly to detection audience.

(R1) Long-range interaction. We select the top-k key features in the entire feature map, and thus the keys are not limited in the local range. We will discuss HoughNet in our revision.

(R2) Thanks for recognizing our "reasonable motivation" and "simple and effective" method. Also thanks for the great suggestion of "truly" integrating representations without a "master". We will keep thinking in this direction.

(R3) Implementation details. The prediction of left-top and right-bottom points are individual. We merge the enhanced feature maps of left-top and right-bottom corner by addition. We will make it clearer in revision.

(R3&R4) Real time cost of the BVR. Table 9 uses an input size of 800×1333 to count the FLOPs. The real inference speed of different models using a V100 GPU (fp32 mode is used) are shown in Table 1. By using a ResNet-50 backbone, the BVR module usually takes less than 10% overhead. By using a larger ResNeXt-101-DCN backbone, the BVR module usually takes less than 3% overhead.

Table 1: Time cost of the BVR module. R-50 and RX-101-D mean ResNet-50 and ResNeXt-101-DCN, respectively.

Method	Backbone	FPS	FPS (+BVR)	Method	Backbone	FPS	FPS (+BVR)
Faster R-CNN	R-50/RX-101-D	21.3/7.5	19.5/7.3	FCOS	R-50/RX-101-D	22.7/7.4	20.7/7.2
RetinaNet	R-50/RX-101-D	18.9/7.0	17.4/6.8	ATSS	R-50/RX-101-D	19.6/7.1	17.9/6.9

(R4) Good performance feels like from putting existing output modalities together; not excited. Actually, this is exactly our goal: while "current object detection systems are doing different representations for objects, this paper wants to integrate them together" (by R2). To achieve this goal, we propose an attention module to bridge these heterogeneous representations. We also propose *novel* techniques of *key sampling* and *shared location embedding* to make the module effective and efficient. Our approach is "simple, effective" (by R2) and general for many different detectors including RetinaNet, FCOS, Faster R-CNN and ATSS (by R1). Although we respect the personal view of "not excited", we would greatly appreciate if the reviewer could also think about "the reasonable motivation" (by R2), "the simple and effective nature" (by R2) and the broad effectiveness (by R1, R2) of this submission.

(R4) Novel CornerNet/CenterNet in an FPN but without details. Thanks for recognizing this novelty, although there are other contributions we value more greatly as described in the response to the last question. The details are in Lines 179-184: we use a CornerNet-style focal loss and assign all ground-truth center/corner points to all FPN levels.

(R4) "ablation experiments on the strongest baseline". Thanks for the suggestion. Noticing BVR achieves 2.0 AP gain on our strongest baseline detector of ATSS (see Table 10), we think whether the proposed technique works well on stronger baselines is well answered. Ablation on the strongest baseline is good, but a more common practice for most papers is to ablate on a reasonably well baseline due to resource limit.

(R4) The centerness branch and center point head. FCOS and ATSS include a classification and a regression branch on the "centerness" points. When applying BVR, we use an independent center point head and corner point head. The center point head plays a different role than the centerness branch: center point is an *auxiliary* representation referring to the box center and aims to strenghthen the *master* features, while centerness is a *master* representation referring to box (center) area and aims for classification and regression. Nevertheless, they could share a same branch, but to strengthen the commonality of the BVR module, we use a same and independent center/corner point heads for all baseline detectors no matter what the master branches are.

(R4) Gains by multi-task learning. Only including an auxiliary point head without using it can boost the RetinaNet baseline by 0.8 AP (from 35.6 to 36.4). Noting the BVR brings a 2.9 AP improvement (from 35.6 to 38.5) under the same settings, the major improvements are not due to multi-task learning. We will add this ablation in our revision.

(R4) A general viewpoint as a contribution. We shall remove it as a contribution in our revision.

(R4) Figures. Figure 1b illustrates typical object representations, while Figure 2 focus on how representation flows throughout the detection pipeline. We will follow the reviewer's suggestions to provide detailed captions for each figure.

(R4) Representation in FCOS. The updated version of FCOS "samples the central portion of ground-truth boxes as positive" (see "ctr. sampling" in Table 3 of FCOS), and we use an implementation of this version in our experiments. We thus categorize it as a center based method. We will make it clearer in our revision.