We sincerely thank the reviewers for their comments. We are pleased to see that our contribution is unanimously found novel the reviewers. The reviewers believe that our experimental results are adequate and convincing. ***R1*** and ***R3*** find that our technical details are clearly explained. We will release the source code and model for reproducibility.

Motivation of our approach: The popular SPP and attention approaches construct the contextual representation that captures the visual relationship between pixels. However, the semantic segmentation task heavily relies on the understanding of the object-level relationship that is ineffectively captured by the pixel-level context. It motivates us to leverage the object regions to compute the regional context. Intuitively, the boundaries of the object regions provide the spatial relationship between the objects. In the same object region, the pixels contain the consistent category information. We resort to these nice properties of the object regions and construct the regional context, which enhances the pixel representations and eventually improves the segmentation performance.

R1-Q1*: The authors should clarify the motivation of the step-by-step* ***RCB*** *and* ***RIB****.*

A: Thanks. The step-by-step **RCB** and **RIB** are motivated by the need for using the regional context to enhance the pixels. We use **RCB** to group the pixels of the image reasonably into different object regions. Based on the boundary and the representative pixels of the region, we construct the spatial and category representations of the object. Next, **RIB** exchange information between object regions, forming the regional context for enhancing the pixels.

R1-Q2*: The authors should explain the extra structure after the backbone in contrast experiments.*

A: In Table 1-3 and 5, we use the backbone ResNet-101 without any extra structure (e.g., SPP). We equip the backbone with **RCB** and **RIB** to produce the contextual representation, which is fed to a 1×1 convolution layer and a softmax layer for segmentation. In Table 4, we use the backbone HRNetV2-W48 equipped with an ASPP structure (see "HRNetV2-W48+ASPP"), along with our approach, to make a fair comparison with the latest OCRNet [33].

R1-Q3*: The authors should optimize the formula.*

A: Thanks. We will optimize the typesetting and reduce the notations of accumulation.

R2-Q1*: Why the RANet capture more context information? The category information in RANet may be not accurate.*

A: In contrast to the SPP or attentional models that capture the pixel-level context, RANet captures the regional context. RANet allows the regional information to be exchanged between the pixels (see **RIB** in Figure 4), forming the pixel representations that contain the pixel-level and regional context. We agree that the category information may be inaccurate. Thus, we use **RCB** to select the representative pixels in different regions, based on the category and boundary information. Our approach produces more reliable context, compared to using the category information alone.

R2-Q2*: How to compute the final segmentation map based on the contextual feature map* O*?*

A: The contextual map O is fed to a 1×1 convolution layer and a softmax layer, for computing the segmentation map.

R2-Q3*: What is the direction of the line in Eq.(1)?*

A: We illustrate an oblique line in Figure 2. Actually, the direction of the line is determined by the locations of the end pixels. Thus, the line can be vertical, horizontal or oblique.

R3-Q1*:* ***RCB*** *and* ***RIB*** *both seem to be very time-consuming.*

A: Please note that **RCB** computes the semantic similarity based on the low-dimensional category score vectors. Though **RIB** needs to select the representative pixels, it effectively reduces the number of pixels that exchange context, and consequently saves the computation. In Table 3, we have shown that our approach can be done at the cost of the reasonable computation, compared to the latest attentional models.

R3-Q2*: The authors should explain the differences between RANet and SPP or other attention mechanisms.*

A: Thanks for your valuable suggestion. Please see "**Motivation of our approach**".

R3-Q3*: Why not use all the pixels in each region? The author needs to give a short explanation.*

A: Thanks. Using many pixels may involve the ambiguous information of the pixels that are near the object boundaries. It degrades the performance (see Table 1). Besides, using the representative pixels saves computation (see Table 3).

R3-Q4*: Why do the representative pixels gradually separate from each other?*

A: We conjecture that the network optimization leads to the separation of the representative pixels to comprehensively represent different contents of the object. We will provide more analyses on the representative pixels in our future work.