We thank all reviewers for their thoughtful comments. Below we integrate all the concerns and clarify them in details.

(R1) Omission comparisons with SOTA methods. As we have claimed in Line 258, we have reported almost all published SOTA methods under a fair comparison in our supplementary material due to the space limit, regardless of feeding images with original or 1.3x enlarged size. Our Beta R-CNN outperforms all of them especially on crowd scenes. We boost SOTA results by $\mathbf{1.1\%}$ $\boldsymbol{MR^{-2}}$ on CrowdHuman, $\mathbf{2.2\%}$ and $\mathbf{5.5\%}$ $\boldsymbol{MR^{-2}}$ on Heavy subset of CityPersons with original and 1.3x enlarged size respectively. Considering the datasets' challenge and other SOTA methods' improvements like [32, 22], our improvement is really convincing. Besides, we thank R1 for raising our concerns about performing our own split on CrowdHuman, it is good advice, and we will take it in the final version.

(R1, R2, R3, R4) Questions about ablation study. Firstly, we thank R3 for the reminding about insufficient analyses of each component in Sec.4.3. Actually we have claimed and validated the effect of each module in Sec.3 like Sec.3.3 for BetaNMS, so we mainly focus on the results to verify their effectiveness in Sec.4.3. We will polish our paper writing more reasonably. **Secondly**, we adopt R2's advice to add a comparative test using softNMS, and the result is $41.1\%\mathrm{MR}^{-2}, 88.0\%\mathrm{AP}$, which is inferior to our method. **Thirdly**, although the improvement of MaskLoss is not as impressive as other modules, but it is proposed to better implement Beta Representation, and actually we have tried other loss functions to supervise Beta Mask but failed to further improve the performance except MaskLoss. We will explore to boost MaskLoss in future work. **Besides**, actually R-CNN models are widely used by most SOTA pedestrian detectors [8,19,22,32]. Comparisons with advanced detectors will be explored in the future work. **Last**, Table 1 clearly shows that although our cascade baseline achieves very high results, our method still achieves **3.5%** $\boldsymbol{MR^{-2}}$ and **3.0%** $\boldsymbol{AP}$ gains, which can verify the effectiveness of our method and well clarify R3's concerns about our improvements.

(R2, R4) Questions about model design. Beta R-CNN is based on Cascade R-CNN by replacing the box regression head with our BetaHead (shown in Fig.3 as Be) to regress eight Beta parameters, which has been introduced in Sec.3.2.1. Besides, BetaMask is proposed to further improve the performance. But due to the coarse results from RPN, BetaMask is hard to contribute to the first BetaHead, thus we add one BetaMask between two adjacent BetaHeads at last.

(R2, R4) Questions about Beta Representation. Beta Representation utilizes 8 parameters to identify a pedestrian, and the 8 parameters are generated from annotated **full-body boxes and visible boxes**. As introduced in Sec 3.1.2, the 8 parameters include 4 boundary parameters and 4 shape parameters. Boundary parameters are consistent with the full-body box, while shape parameters control the peak and width of 2D beta distribution. For human visual habits, we tend to emphasize the visible parts, so we assign different weights for visible/non-visible parts and calculate the mean μ and variance σ. Based on μ and σ, we could deduce the shape parameters according to the beta distribution properties. Then, the 2D beta distribution could highlight the visible part and suppress the non-visible part.

(R2, R4) Comparisons with related works. Due to space limit, our descriptions of related works in Sec.2 maybe a bit concise, we will add more details in the final version. Besides, R4 mentioned $Object\ as\ Distribution$ [27]. The similarity between [27] and our method is that we both adopt distribution to represent objects, but actually there are many differences. [27] utilizes the bivariate normal distribution and we have claimed its weaknesses in Line 94-97. [27] can't model the unpredictable visible patterns about pedestrians. Also, [27] needs more complex training strategies and its performance is considerably poor. As for KL, i.e., Kullback-Leibler divergence, it is a common metric to measure the distance between two distributions, and we adapt it to our BetaNMS.

(R3) Speed/accuracy trade-off. Each proposed component in Beta R-CNN is light-weight with little computation cost. We take CrowdHuman validation set with 800x1400 input size to conduct speed experiments on NVIDIA 2080Ti GPU with 8 GPUs, and the average speeds are **0.483s/image (Cascade R-CNN baseline)** and **0.487s/image (Beta R-CNN)** respectively. Besides, the total flops of them are 52.77G and 52.78G respectively. The difference can be negligible.

(R4) Is Beta R-CNN practical? Firstly, the proposed components like BetaHead in Beta R-CNN are all optimization-friendly, so the training of Beta R-CNN is as easy as Faster R-CNN and Cascade R-CNN. **Secondly**, Beta Representation is based on beta distribution, it is intuitive to adopt BetaNMS (based on KL) to measure distance between distributions, which achieves further performance improvement comparing with other NMS strategies without any extra cost. **Thirdly**, Beta Representation, as well as BetaHead, BetaMask, BetaNMS are all flexible enough to be integrated into other two-stage or single-shot detectors, and are also compatible with existing optimization methods like EMD[32] to further boost their performance. **Last**, Beta R-CNN is proposed for occlusion and crowd issues but it still works well on standard scenes like Reasonable subset of CityPersons. So Beta R-CNN can also be generalized on Caltech.

(R4) Questions about evaluation metric. Some connection exists between MR^{-2} and AP, but higher AP doesn't mean lower MR^{-2}. MR^{-2} is more suitable by emphasizing FP&FN beyond AP, which is critical in pedestrian detection.

(R1, R2, R3, R4) Paper writing. We really appreciate all the useful suggestions about our paper writing, we will polish our paper in the final version to fix all of them, i.e. figure out a better title, improve captions of tables and figures (R1), refine Fig.4 (R3), clarify notations and definitions in details(R1,R3), and other common written problems.