We thank the reviewers for the insightful comments, and we will update the paper accordingly.

R1Q1: compare with uncertainty learning methods. **A:** Using the loss from ["What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?", NeurIPS 2017] on UCF-QNRF we obtain 103.2/168.2 (MAE/MSE), which is worse than our proposed loss. We will add more references about uncertainty learning.

R1Q2, R3Q2: Why Φ and Ψ are approximated with Gaussian? **A:** Using Gaussians for approximate inference is common, since they are tractable and can be estimated from 1st and 2nd moments. Extensions of the central limit theorem prove that sums of independent non-identical r.v.s converge to Gaussian. Indeed, in Fig 2c, the distribution is tending to Gaussian with just 3 annotations, and we observe this tendency becomes stronger with more annotations. We have also tried Gamma distributions for the approximation, but the results are worse (MAE 89.7 UCF-QNRF).

R1Q3: Tab. 2 should include "L2+Reg (L_i)". **A:** The result of L2+Reg is 94.5/160.0, which is worse than our loss.

R1Q4: The effect of α when β is large? Robustness of L2 to different β. **A:** We evaluate the effect of α when $\beta = 16$ in Fig. R1, and the proposed loss is effective for a larger β. We also show L2 loss with large β for different noise levels in Fig. R2. For large β, the performance is bad because the density map is over-smoothed.

R1Q5, R3Q5, R4: Try on other tasks to show generalizability? **A:** We propose a general framework to model noise in point-wise annotations that are converted to response maps. In future work, we will investigate other tasks that use response maps, such as human joint detection and visual object tracking. However, we think the derivation of the proposed framework and approximation method, as well as extensive experiments on crowd counting, are strong enough for a standalone paper that can inspire others.

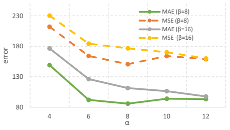

Fig. R1: large β.

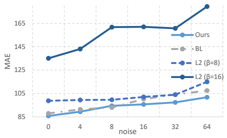

Fig. R2: Annot. noise.

R2Q1: How correlation between pixels is modeled? **A:** Each dimension of the m.v. Gaussian corresponds to one pixel in the density map, and thus the covariance matrix models the correlations between pixels (see L157-169, Figs. 2 & 3). Equivalently, the loss is Mahalanobis distance (Eq. 10).

R2Q2: Other backbones CSRNet and MCNN? **A:** The performance of different backbones is reported in Table 1 (Sec. 4.2.1). Since VGG19 works better in the ablation, we only show VGG19 in the remaining experiments.

R2Q3: Integrate proposed loss with existing SOTA crowd counting? **A:** VGG19 is the strongest backbone before 2020, and we evaluate other backbones in Table 1. We also use our loss function with self-correlation strategy ["Adaptive Dilated Network with Self-Correction Supervision for Counting", CVPR 2020] (*published after the NeurIPS submission deadline*), and achieve better performance on UCF-QNRF (84.3/142.9 vs. 85.8/150.6).

R2Q4: Missing discussion between related and proposed works. **A:** See L82-84 and L93-98 for these discussions.

R2Q5: grammatical errors and typos. **A:** We will revise the paper.

R2Q6: Explain L119-123. **A:** i.i.d Gaussian noise (L2 norm) assumes independent noise between pixels. However, a noisy annotation actually induces correlated noise between pixels in the density map, see R2Q7 for details.

R2Q7: Explain L26. **A:** If noise is added to the annotation, then the density map values in nearby pixels change in a common way (i.e., correlated). For example, in Fig 2a, if the right-most green dot moves towards $x^{(0)}$, then the density values at $x^{(0)}$ and $x^{(1)}$ will both increase. Also, if it moves away, then density values at $x^{(0)}$,$x^{(1)}$ will both decrease.

R3Q1: Does not handle missing or duplicate annotations. **A:** Missing annotations are handled by the background model (L197-203), which adds a "virtual dot" close to each pixel (equivalent to a hypothetical missing annotation). Duplicate annotations are rare, and will be corrected by the annotator in the review phase. Modeling displacement noise already yields substantial benefits, and we will investigate other kinds of annotation noise in future work.

R3Q3: Why $\eta_{\mathbf{r}_i}$ is non-central χ^2? **A:** $\eta_{\mathbf{r}_i}$ is a m.v. Gaussian with mean $\mathbf{r}_i$ and identity covariance. $\|\eta_{\mathbf{r}_i}\|^2$ is the sum squares of Gaussian r.v.s with non-zero mean and unit variance, and thus a non-central χ^2 distribution (by definition).

R3Q4: Small shifts in annotations will not affect performance much, since NN can be robust if receptive size of the network is large enough. **A:** Shifts in annotations cause changes in the GT *target output*, and it is more appropriate to modify the loss function in order to make the network less sensitive to these output changes (see Figs. 4 & 7). On the other side, using large receptive field and pooling layers will make the network invariant to local shifts in the *input image*, but this does not make it better at handling ambiguous outputs.

R4Q1: Novelty is limited; modeling annotation noise explored in [4]. **A:** Our approach is quite different from [4]. [4] uses the annotations to compute weights on the density map pixels, whereas our approach is a generative model mapping annotation noise to density map noise. [4] handles each annotation separately, while our work models correlations induced by multiple annotations (L148-149).