Reviewer 1. **Q1:** *The paper does not address if this is of interest to broader vision/ml community.* **A1:** Crowd counting is a well established research problem in Computer Vision, with more than four dozen papers published at top-tier conferences in the last two years, including the 12 papers mentioned by R1. Our paper is motivated by crowd counting applications, but it actually addresses the underlying fundamental problem: spatial density estimation, which has broader interest to machine learning and other communities. We believe that crowd counting is a very appropriate and salient evaluation domain for spatial density estimation, especially in the age of social distancing. In addition, almost all papers including those listed by R1 used the Gaussian smoothed Ground Truth (GGT), so we believe that a theoretical analysis of GGT is necessary and that NeurIPS is a proper venue to present our theoretical results.

Q2: *Missing discussions of several recent works, especially [2,9,10] which focused on improving representation of ground-truth for training the networks.* **A2:** We would be happy to cite these papers in our revised paper. Notably, it is neither possible nor necessary to cite every single crowd counting paper given the large number of papers. We could only discuss the major approaches and representative papers. All the listed papers except [5] used the GGT, which is discussed in lines 38–41. Refs [2,9,10] belong to the group of papers that use adaptive kernel widths, discussed in lines 82-84. Ref [5] belongs to the detection-then-count approach, line 73–76.

Q3: *Missing comparisons to recent methods that have better/comparable results. The improvements are insignificant.*

A3: Thanks for the references, and this table compares with methods in the references. Considering all four datasets as a whole, our method outperforms the other methods. Considering individual datasets, QNRF is the most difficult and largest one, and the performance gaps between our method and others are wide. For ShTech datasets, our method performed comparable to the bests. On UCF-CC 50, the comparison should be taken with a grain of salt due to small number of images (50) and different ways of data splitting for five-fold cross validation. Moreover, none of these methods were evaluated on the large-scale NWPU dataset. Our method ranked first in the leaderboard at the time of submission, reducing SOTA error from 105 to 88.

	QNRF		ShTech A		ShTech B		CC-50	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
[2]	97.5	165.2	60.2	**94.1**	8.0	15.5	233	300
[11]	99.1	159.2	60.6	96	**6.8**	**10.3**	216	302
[12]	111	190	59.4	102	7.9	12.9	239	319
[13]	96.5	170.2	**59.2**	98.1	8.1	12.2	**185**	278
[14]	118	180.4	62.9	**94.9**	8.1	13.4	256	348
CAN	107	183	62.3	100.0	7.8	12.2	212	**243**
BL	88.7	154.8	62.8	101.8	7.7	12.7	229	308
Ours	**85.6**	**148.3**	**59.7**	**95.7**	**7.4**	**11.8**	211	291

Reviewer 2. **Q4:** *DM-Count is worse than CAN in RMSE on UCF-CC-50 and NWPU.* **A4:** One reason is that CAN is trained by minimizing the MSE loss, while DM-Count optimizes for the MAE loss. On NWPU, the RMSEs are close (388.6 vs 386.5). The comparison on UCF-CC-50 should be taken with a grain of salt, as explained in Answer A3.

Q5: *Why OT loss approximate well for dense regions, but poorer for the sparse regions.* **A5:** Dense areas have higher probability values than sparse areas, and in general more probability masses must be transported between dense areas. The OT loss is therefore dominated by loss from dense areas. The dense areas have higher priority in the optimization, especially when using a finite number of Sinkhorn iterations for computing the OT distance.

Q6: *Without using the TV loss, DM-Count performed worse than Bayesian Loss (BL).* **A6:** This is due to the pre-mature stopping of the Sinkhorn algorithm. Tab. 3 reports the performance with 100 Sinkhorn iterations. Using 200 Sinkhorn iterations, OT loss + Counting loss outperforms BL with MAE 85 and MSE 154. As the TV loss treats both dense and sparse areas equally, sparse areas can be optimized well with TV loss and fewer Sinkhorn iterations (100).

Q7: *Is this possible to transfer the DM-Count to other similar tasks, like keypoint regression?* **A7:** Thanks for this interesting suggestion. We think it may be possible, given that the GT for keypoint estimation is also a dot map.

Reviewer 3. **Q8:** *The effects of the hyper-parameters should be studied.* **A8:** On QNRF, by fixing λ_2 to 0.01 and tuning λ_1 from 0.01, 0.05 to 0.1, the MAE varies from 87.2, 86.2 to 85.6. By fixing λ_1 to 0.1 and changing λ_2 from 0.01, 0.05 to 0.1, the MAE varies from 85.6, 87.8 to 88.5. See also A5, A6 for the effects of the OT loss and TV loss.

Q9: *"Total variation should be clearly defined and explained. The derivation of Eq. 6 should be further explained."* **A9:** We will clarify that in the context of training loss, Total Variation refers to the total variation distance of two probability measures, not the total variation of a function. A formal definition can be found in Definition 2.4, pg 83, Tsybakov: Introduction to Nonparametric Estimation. Eq. 6 in our paper is Lemma 2.1, pg 84. We will cite and clarify.

Q10: *What is the maximum Sinkhorn iterations in the experiments? Will these iterations significantly slow down the training speed?* **A10:** We experiment with different numbers of iterations in Tab 1 of the supplementary material. The performance plateaus after 100 iterations. In our experiments, we used a maximum of 100 iterations, and the OT part takes 30% of the total training time. Thus the OT part does not significantly slow down the training speed.

Q11: *Reproducibility* **A11:** Implementation details are provided in Sec 2.3 of the supplementary material.

Reviewer 4. **Q12:** *Missing ablation studies of using single loss: only OT loss, only TV loss.* **A12:** the OT and TV losses take normalized density maps (probability distributions) as inputs (Eqs. 4 and 6), so it is not possible to obtain the absolute count using either loss by itself. The counting loss is always needed. In Tab. 3, we showed the performance of counting loss + OT loss (or TV loss). Besides, we will report the hyper-parameter study (see Answers A8).

Q13: *Quantitative performance in high-density region.* **A13:** The MAEs in high density images (Ground truth count over 1000 in QNRF test set) are 211 (DM-Count), 238 (BL) and 311 (pixel-wise loss). DM-Count outperforms two baselines significantly for high density images. We will address other minor issues and release code upon acceptance.