Supplemental for A Generalized Loss Function for Crowd Counting and Localization

Jia Wan Ziquan Liu Antoni B. Chan
Department of Computer Science, City University of Hong Kong
jiawan1998@gmail.com, ziquanliu2-c@my.cityu.edu.hk, abchan@cityu.edu.hk

A. Gradient of counting loss in DM-Count

The counting loss is

$$\mathcal{L}_C = |\mathbf{1}^T\boldsymbol{z} - \mathbf{1}^T\boldsymbol{a}| = \begin{cases}\mathbf{1}^T\boldsymbol{z} - \mathbf{1}^T\boldsymbol{a}, & \mathbf{1}^T\boldsymbol{z} > \mathbf{1}^T\boldsymbol{a},\\ \mathbf{1}^T\boldsymbol{a} - \mathbf{1}^T\boldsymbol{z}, & \mathbf{1}^T\boldsymbol{z} < \mathbf{1}^T\boldsymbol{a}.\end{cases} \tag{1}$$

where $\boldsymbol{z}$ is the GT dot annotation map, $\boldsymbol{a}$ is the predicted density map, and $\mathbf{1}$ is the vector of n ones.

Taking the derivative w.r.t. the prediction $\boldsymbol{a}$ yields

$$\frac{d\mathcal{L}_C}{d\boldsymbol{a}} = \begin{cases}-\mathbf{1}, & \mathbf{1}^T\boldsymbol{z} > \mathbf{1}^T\boldsymbol{a},\\ \mathbf{1}, & \mathbf{1}^T\boldsymbol{z} < \mathbf{1}^T\boldsymbol{a}.\end{cases} \tag{2}$$

Thus, the gradient signal to decrease the count loss is to increase/decrease the values of *all* pixels by the same value.

B. Background model for Bayesian Loss

The background model used in BL [1] is

$$\mathcal{L}_{BG} = |0 - \sum_{i=1}^{n}\bar{\omega}_i a_i| = \sum_{i=1}^{n}\bar{\omega}_i a_i, \tag{3}$$

since both $\bar{\omega}$ and a_i are positive. The weight is

$$\bar{\omega}_i = \frac{\bar{k}_i}{\bar{k}_i + \sum_{j=1}^{m} K_{ij}}, \tag{4}$$

where $\bar{k}_i$ is the Gibbs kernel for a "dummy" background point $\bar{\boldsymbol{y}}_i$ given by

$$\bar{\boldsymbol{y}}_i = \boldsymbol{y}_{\eta(i)} + d\frac{\boldsymbol{x}_i - \boldsymbol{y}_{\eta(i)}}{||\boldsymbol{x}_i - \boldsymbol{y}_{\eta(i)}||}, \tag{5}$$

$$\bar{k}_i = \exp(-||\boldsymbol{x}_i - \bar{\boldsymbol{y}}_i||^2/\varepsilon), \tag{6}$$

where $\boldsymbol{y}_{\eta(i)}$ is the nearest annotation to $\boldsymbol{x}_i$, and d is a hyper-parameter. [1] shows $\bar{k}_i$ can be written as

$$\bar{k}_i = \exp(-(d - ||\boldsymbol{x}_i - \boldsymbol{y}_{\eta(i)}||)^2/\varepsilon) \tag{7}$$

We can rewrite $\bar{k}_i$ as

$$\bar{k}_i = \exp(-\frac{1}{\varepsilon}(d^2 - 2d||\boldsymbol{x}_i - \boldsymbol{y}_{\eta(i)}|| + ||\boldsymbol{x}_i - \boldsymbol{y}_{\eta(i)}||^2)) \tag{8}$$

$$= \exp(\frac{2d}{\varepsilon}||\boldsymbol{x}_i - \boldsymbol{y}_{\eta(i)}|| - \frac{d^2}{\varepsilon})\exp(-\frac{1}{\varepsilon}||\boldsymbol{x}_i - \boldsymbol{y}_{\eta(i)}||^2) \tag{9}$$

$$= \exp(\frac{2d}{\varepsilon}||\boldsymbol{x}_i - \boldsymbol{y}_{\eta(i)}|| - \frac{d^2}{\varepsilon})K_{i,\eta(i)} \tag{10}$$

Thus the weight $\hat{\omega}$ becomes

$$\bar{\omega}_i = \exp(\frac{2d}{\varepsilon}||\boldsymbol{x}_i - \boldsymbol{y}_{\eta(i)}|| - \frac{d^2}{\varepsilon})\bar{\pi}_i, \tag{11}$$

where $\bar{\pi}_i$ is a weight computed from the distance to the nearest annotation,

$$\bar{\pi}_i = \frac{K_{i,\eta(i)}}{\bar{k}_i + \sum_{j=1}^{m} K_{ij}}. \tag{12}$$

C. Full-Sized Figures

Here we show the full-size figures for Figures 5 and 7 in the paper.

D. Localization Comparison

In Fig. 8, we show a comparison of the localization results for different loss functions.

E. More Examples of localization results on NWPU-test set

We further show more examples of localization result on NWPU-test set.

References

[1] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6142–6151, 2019. 1

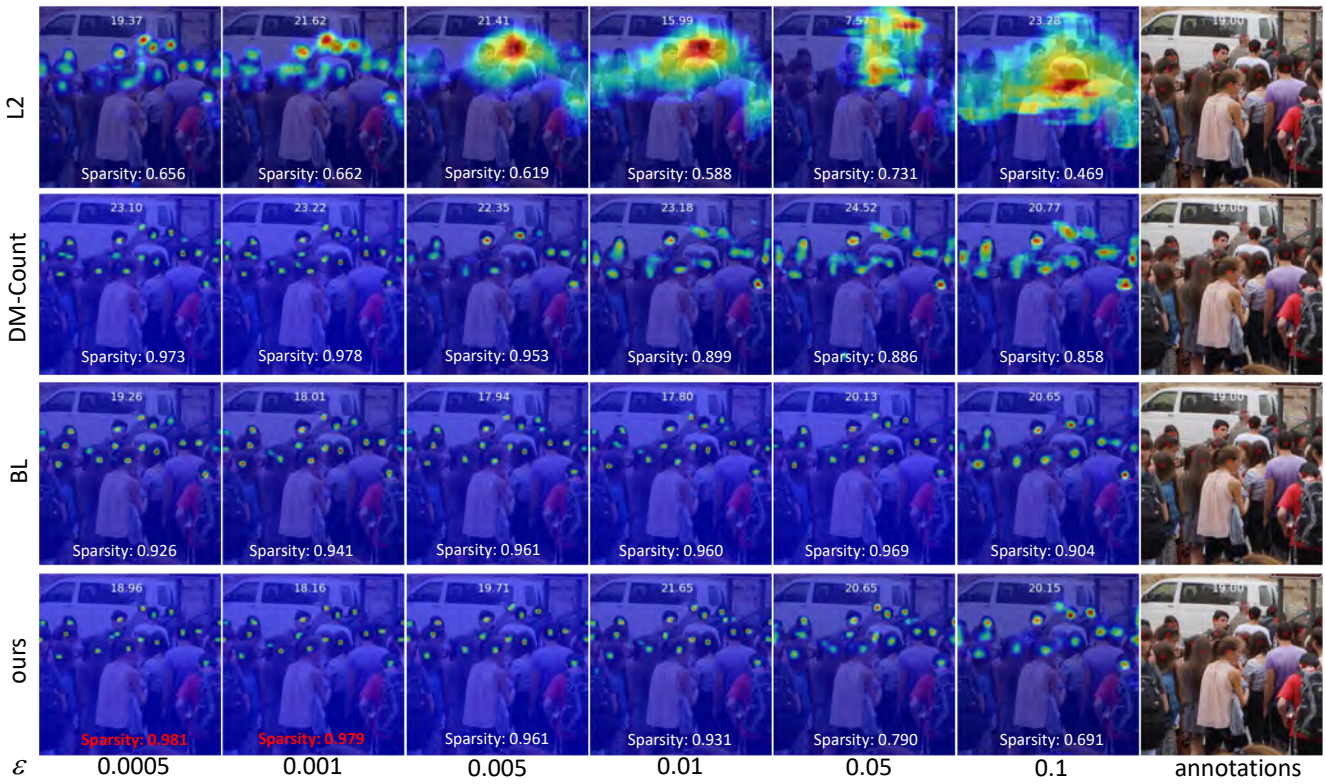
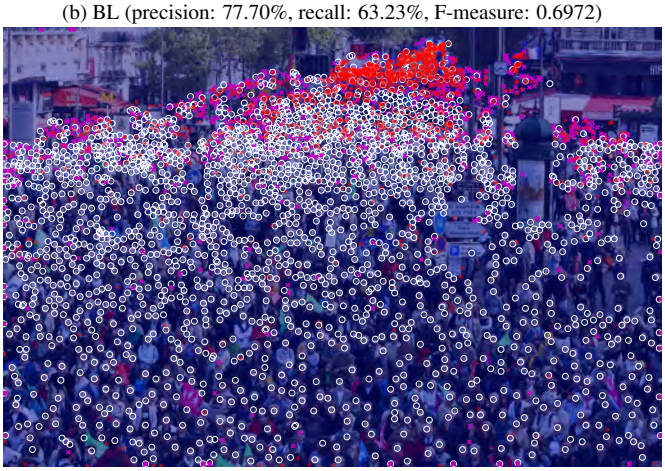


Figure 5: Visualization of density maps predicted from models trained with different loss functions and blur factors. Note that the width and height of the trained images patches are normalized to 1. The sparsity is defined as the percentage of pixels with density less than 0.001 and the two most sparse density maps are shown in red bold.

Figure 7: Example localization result on NWPU-Crowd test set. We propose a generalized loss function based on unbalanced optimal transport theory for learning crowd density maps, which can be used for both crowd counting and localization. Our method achieves the best performance on NWPU-Crowd benchmark for both tasks.

(a) L2 (precision: 91.49%, recall: 56.99%, F-measure: 0.7023)

(b) BL (precision: 77.70%, recall: 63.23%, F-measure: 0.6972)



(c) DM-Count (precision: 61.93%, recall: 73.18%, F-measure: 0.6709)

(d) ours (precision: 90.17%, recall: 70.87%, F-measure: 0.7936)

Figure 8: Visualization of localization for different loss functions. White circles are true positives, red dots indicate false negatives, and magenta crosses are false positives. L2 generates smooth density map and thus many small objects are not detected. DM-Count has many duplicate detections, where red dots and magenta crosses approximately overlapped since it is easy to over-fit with dot maps for pixel-wise supervision. The background model used in BL tends to generate blurry density maps for high-density regions, which results in worse precision and recall compare to the proposed loss.

Figure 9: Examples of localization result on NWPU-Crowd test set.