Supplemental for Robust Zero-Shot Crowd Counting and Localization With Adaptive Resolution SAM

Jia Wan[1], Qiangqiang Wu[2], Wei Lin[2], and Antoni Chan[2]

[1] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen

[2] Department of Computer Science, City University of Hong Kong

jiawan1998@gmail.com, qiangqwu2-c@my.cityu.edu.hk, elonlin24@gmail.com, abchan@cityu.edu.hk

1 Gaussian Mixture Model for Pseudo Point Labels

A soft mask $M \in \mathbb{R}^{h\times w}$ generated via SEEM can also be represent as $M = \{(s_i,\ x_i)\}_{i=1}^{h\times w}$, where s_i is the score value locating at x_i. To find the pseudo point label indicating the human head, we use a mixture of two Gaussian distributions to fit the mask M:

$$G = p(x) = \sum_{j=1}^{2} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j), \tag{1}$$

in which these parameters are estimated effectively through the Expectation Maximization (EM) algorithm in practice.

In the *E-step*, the soft assignments are computed according to the current estimated G. In particular, the likelihood that assigns the i-th score $(s_i,\ x_i)$ to the j-th Gaussian distribution is formulated as:

$$\hat{z}_{ij} = p(z_i = j|x_i, G) = \frac{\pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}{\sum_{k=1}^{2} \pi_k \mathcal{N}(x_i|\mu_j, \Sigma_k)}. \tag{2}$$

After all $\hat{z}_{ij}$ is obtained, the parameters in the two-Gaussian mixture G is updated in *M-step* by maximizing the likelihood:

$$\hat{N}_j = \sum_{i=1}^{h\times w} s_i \hat{z}_{ij}, \tag{3}$$

$$\hat{\pi}_j = \frac{\hat{N}_j}{\sum_{i=1}^{h\times w} s_i}, \tag{4}$$

$$\hat{\mu}_j = \frac{1}{\hat{N}_j} \sum_{i=1}^{h\times w} s_i \hat{z}_{ij} x_i, \tag{5}$$

$$\hat{\Sigma}_j = \frac{1}{\hat{N}_j} \sum_{i=1}^{h\times w} s_i \hat{z}_{ij} (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^\top. \tag{6}$$

With the estimated parameters, we denote the mean $\hat{\mu}_j$ of the Gaussian component with the smaller vertical coordinate (height) as the head location.

2 Performance on NWPU-Crowd dataset

We compare our performance with existing supervised methods on the NWPU-Crowd test set for reference. The result is shown in Table 1. The proposed method achieves comparable performance to some supervised models.

Table 1: Comparison with supervised methods on NWPU (test).

Method	Label	MAE ↓	MSE ↓	Prec	Rec	F1
TinyFaces	Point	272.4	764.9	0.529	0.611	0.567
MCNN	Point	232.5	714.6	-	-	-
SANet	Point	190.6	491.4	-	-	-
GeneralizedLoss	Point	79.3	346.1	0.800	0.562	0.660
Ours	None	168.4	547.5	0.762	0.510	0.611