Point-to-Region Loss for Semi-Supervised Point-Based Crowd Counting

Supplementary Material

Algorithm 1: P2R in fully-supervised training

Hyperparameter: μ, τ, and λ

/* Score map predicted by model: */

Input: $\mathcal{P} = \{\boldsymbol{p} \in \mathbb{R}^n, \boldsymbol{x} \in \mathbb{R}^{n\times 2}\}$

/* The coordinates of ground truth points: */

Input: $\boldsymbol{x}' \in \mathbb{R}^{m\times 2}$

Output: loss value $\mathcal{L}_l$.

/* Compute the l_2 distance matrix $\mathbf{L}_2$ */

$\mathbf{L}_2 \in \mathbb{R}^{n\times m}$: $\mathbf{L}_{2[i,\,j]} \leftarrow \|\boldsymbol{x}_{[i]} - \boldsymbol{x}'_{[j]}\|_2$

/* Obtain the minimum $\boldsymbol{d}$ and the corresponding column index $\boldsymbol{k}$ of each row in $\mathbf{L}_2$ */

Initialize $\boldsymbol{d} \in \mathbb{R}^n$ and $\boldsymbol{k} \in \mathbb{N}^n$

for $i \leftarrow 1$ *to* n **do**

  $\boldsymbol{d}_{[i]} \leftarrow \min_j \mathbf{L}_{2[i,\,j]}$

  $\boldsymbol{k}_{[i]} \leftarrow \arg\min_j \mathbf{L}_{2[i,\,j]}$

/* Compute the P2R matching matrix $\mathbf{M}$ in (18) */

$\mathbf{M}_f \leftarrow \mathbf{0}_{n\times m}, \quad \boldsymbol{\beta} \leftarrow \mathbf{0}_n$

for $i \leftarrow 1$ *to* n **do**

  $\mathbf{M}_{f\left[i,\,\boldsymbol{k}_{[i]}\right]} \leftarrow 1, \quad \boldsymbol{\beta}_{[i]} \leftarrow (\boldsymbol{d}_{[i]} < \mu)$

$\mathbf{M} \leftarrow \mathbf{M}_f \odot (\boldsymbol{\beta}^\top \mathbf{1}_m)$

/* Compute the cost $\mathbf{C}$ in (21) for foreground pixel selection, in which the subsection in (21) can be implemented by element-wise division */

$\mathbf{C} \in \mathbb{R}^{n\times m}$: $\mathbf{C}_{[i,\,j]} \leftarrow \tau\mathbf{L}_{2[i,\,j]} + \log\left(\frac{1}{\boldsymbol{p}_{[i]}} - 1\right)$

$\mathbf{C} \leftarrow \mathbf{C} \oplus \mathbf{M}$

/* Estimate the learning objective $\hat{\boldsymbol{p}}$ by marking the minimum in each column of $\mathbf{C}$ using $\hat{\mathbf{M}}$ */

$\hat{\mathbf{M}} \leftarrow \mathbf{0}_{n\times m}$ **for** $j \leftarrow 1$ *to* m **do**

  $k \leftarrow \arg\min_i \mathbf{C}_{[i,\,j]}$

  $\hat{\mathbf{M}}_{[k,\,j]} \leftarrow 1$

$\hat{\boldsymbol{p}} \leftarrow \hat{\mathbf{M}}^\top \mathbf{1}_m$

/* compute the weighted binary cross entropy */

$\mathcal{L}_l \leftarrow -\lambda\hat{\boldsymbol{p}}^\top \log \boldsymbol{p} - (\mathbf{1}_n - \hat{\boldsymbol{p}})^\top \log(\mathbf{1}_n - \boldsymbol{p})$

return $\mathcal{L}_l$

Algorithm 2: P2R in semi-supervised training

Hyperparameter: μ, τ, η, and λ

/* prediction of student: */

Input: $\mathcal{P}_s = \{\boldsymbol{p} \in \mathbb{R}^n, \boldsymbol{x} \in \mathbb{R}^{n\times 2}\}$

/* pseudo-labels generated by teacher: */

Input: $\mathcal{P}'_t = \{\boldsymbol{p}'_t \in \mathbb{R}^m, \boldsymbol{x}'_t \in \mathbb{R}^{m\times 2}\}$

Output: loss value $\mathcal{L}_u$.

/* Compute the l_2 distance matrix $\mathbf{L}_2$ */

$\mathbf{L}_2 \in \mathbb{R}^{n\times m}$: $\mathbf{L}_{2[i,\,j]} \leftarrow \|\boldsymbol{x}_{s[i]} - \boldsymbol{x}'_{t[j]}\|_2$

/* Obtain the minimum $\boldsymbol{d}$ and the corresponding column index $\boldsymbol{k}$ of each row in $\mathbf{L}_2$ */

Initialize $\boldsymbol{d} \in \mathbb{R}^n$ and $\boldsymbol{k} \in \mathbb{N}^n$

for $i \leftarrow 1$ *to* n **do**

  $\boldsymbol{d}_{[i]} \leftarrow \min_j \mathbf{L}_{2[i,\,j]}$

  $\boldsymbol{k}_{[i]} \leftarrow \arg\min_j \mathbf{L}_{2[i,\,j]}$

/* Compute the P2R matching matrix $\mathbf{M}_{st}$ */

$\mathbf{M}_f \leftarrow \mathbf{0}_{n\times m}, \quad \boldsymbol{\beta} \leftarrow \mathbf{0}_n$

for $i \leftarrow 1$ *to* n **do**

  $\mathbf{M}_{f\left[i,\,\boldsymbol{k}_{[i]}\right]} \leftarrow 1, \quad \boldsymbol{\beta}_{[i]} \leftarrow (\boldsymbol{d}_{[i]} < \mu)$

$\mathbf{M}_{st} \leftarrow \mathbf{M}_f \odot (\boldsymbol{\beta}^\top \mathbf{1}_m)$

/* Compute the cost $\mathbf{C}$ in (21) for foreground pixel selection, in which the subsection in (21) can be implemented by element-wise division */

$\mathbf{C} \in \mathbb{R}^{n\times m}$: $\mathbf{C}_{[i,\,j]} \leftarrow \tau\mathbf{L}_{2[i,\,j]} + \log\left(\frac{1}{\boldsymbol{p}_{[i]}} - 1\right)$

$\mathbf{C} \leftarrow \mathbf{C} \oplus \mathbf{M}_{st}$

/* Estimate the learning objective $\hat{\boldsymbol{p}}$ by marking the minimum in each column of $\mathbf{C}$ using $\hat{\mathbf{M}}$ */

$\hat{\mathbf{M}} \leftarrow \mathbf{0}_{n\times m}$ **for** $j \leftarrow 1$ *to* m **do**

  $k \leftarrow \arg\min_i \mathbf{C}_{[i,\,j]}$

  $\hat{\mathbf{M}}_{[k,\,j]} \leftarrow 1$

$\hat{\boldsymbol{p}} \leftarrow \hat{\mathbf{M}}^\top \mathbf{1}_m$

/* Compute the confidence diagonal matrix $\mathbf{Z}$ */

$\boldsymbol{\xi} \in \{0, 1\}^m$: $\boldsymbol{\xi}_{[i]} \leftarrow \boldsymbol{p}'_{t[i]} > \eta$

$\mathbf{Z} \leftarrow \mathbf{0}_{n\times n}$

for $i \leftarrow 1$ *to* n **do**

  $\mathbf{Z}_{[i,\,i]} \leftarrow \mathbf{M}_{st[i,\,:]}^\top \boldsymbol{\xi} + (1 - \boldsymbol{\beta}_{[i]})$

/* compute the weighted binary cross entropy */

$\mathcal{L}_u \leftarrow -\lambda\hat{\boldsymbol{p}}^\top \mathbf{Z} \log \boldsymbol{p} - (\mathbf{1}_n - \hat{\boldsymbol{p}})^\top \mathbf{Z} \log(\mathbf{1}_n - \boldsymbol{p})$

return $\mathcal{L}_u$

In this supplemental material, we present the following contents:

- Sec. 8: Pseudo-code of P2R for training.
- Sec. 9: Implementation details of P2R.
- Sec. 10: Ablation study on $\mathcal{S}(\cdot)$ *vs.* $I(\cdot)$ in (4).
- Sec. 11: Proof of the unlabeled loss (9).
- Sec. 12: Comparison of P2R with state-of-the-art loss functions in fully-supervised crowd counting.
- Sec. 13: Visualization of P2R predictions on UCF-QNRF dataset images.

8. Pseudo Code of P2R

In Algo. 1 and Algo. 2, we present the pseudo-code to compute the loss for data with ground truth (GT) labels and pseudo-labels, respectively. In Algo. 2, the parts that dif-

Method	Identity		Inv-Sigmoid	
	MAE	MSE	MAE	MSE
P2P [57]	52.74	85.60	**52.50**	**82.94**
P2R (ours)	53.30	83.01	**51.02**	**79.68**

Table 3. Ablation study on $\mathcal{S}(\cdot)$.

fer from those in Algo. 1 are highlighted in purple. The comparison between these two algorithms shows that most of the steps are the same, except for the involvement of confidence computation in Algo. 2.

Also note that the computation of P2R is efficient since the loops in Algorithm 1 and Algorithm 2 can be written as matrix operations, which can be executed in parallel on a GPU.

9. Implementation Details

Data pre-processing. Images in the datasets are cropped into 256×256 for training. For labeled images, we apply horizontal flips to each cropped sample with a probability of 0.5 and randomly resize the images with a scale factor between 0.7 and 1.3. For unlabeled data, the set of weak data augmentation operations is the same as those for labeled data, while strong data augmentation includes adjustments of brightness, contrast, saturation, and hue, conversion from color images to grayscale images, the addition of Gaussian blur with different kernel sizes, and cutout. Besides, Cutout is also implemented, as displayed in Fig. 1.

Training process. In all experiments, we train the model for 1500 epochs with a batch size of 16. Only labeled data are used in the first 100 epochs for initialization. After that, α in (17) is gradually increased from 0 to 1 with a step of 0.01. Adam [**?**] serves as the optimizer, with a learning rate of 5×10^{-5} for the decoder $\mathcal{D}$ and 1×10^{-5} for the backbone $\mathcal{F}$. Furthermore, the loss in the cut-out patch is directly set to 0 to avoid unreasonable optimization.

10. Ablation Study on Cost Function

In (4), we use the inverse sigmoid function,

$$\mathcal{S}(p) = -\log\left(\frac{1}{p} - 1\right), \tag{24}$$

rather than the identity operator used in vanilla P2PNet [57] for better performance. We present the empirical results in Table 3 to demonstrate its advantage. It shows that the inverse sigmoid function performs better than the identity operator on both P2P [57] and P2R.

11. The Proof of Ill-Posed Unlabeled Loss

Under the P2P framework, we demonstrate that the loss function for unlabeled data, formulated as (8), is ill-posed,

Loss function	Point-based counting model	ShTech B		QNRF		FPS
		MAE	MSE	MAE	MSE	
L2 [73]	✗	7.6	13.0	102.0	171.4	**1503.8**
BL [38]	✗	7.7	12.7	88.7	154.8	274.22
GL [61]	✗	7.3	11.7	84.3	147.5	26.18
DMC [63]	✗	7.4	11.8	85.6	148.3	25.49
P2P [57]	✓	6.3	9.9	85.3	154.5	2.32
P2R (ours)	✓	**6.2**	**9.8**	**83.3**	**138.1**	156.25

Table 4. Comparison of counting losses (100% Label Pct.)

since the second term for the background part is set to 0, as shown in (9). This is easy to follow because

$$\mathbf{Z}\hat{\boldsymbol{p}}_t = \mathrm{diag}(\mathbf{M}_{st}\boldsymbol{\zeta})(\mathbf{M}_{st}\mathbf{1}_n) \tag{25}$$

$$= (\mathbf{M}_{st}\boldsymbol{\zeta}) \odot (\mathbf{M}_{st}\mathbf{1}_n) = \mathbf{M}_{st}\boldsymbol{\zeta}. \tag{26}$$

Equation in (26) holds since each row of $\mathbf{M}_{st}$ is an all-zero or one-hot vector since it is matrix result of bipartite-graphs matching between prediction and GT. Taking any row of $\mathbf{M}_{st}$ and denoting it as $\boldsymbol{m} \in \mathbb{R}^m$, we have:

$$\boldsymbol{m}^\top\boldsymbol{\zeta} = \begin{cases} 1, & \text{if } \exists j : \boldsymbol{m}_{[j]} = \boldsymbol{\zeta}_{[j]} = 1 \\ 0, & \text{otherwise} \end{cases}, \tag{27}$$

$$\boldsymbol{m}^\top\mathbf{1}_m = \begin{cases} 1, & \text{if } \exists j : \boldsymbol{m}_{[j]} = 1 \\ 0, & \text{otherwise} \end{cases}, \tag{28}$$

which can be combined to result in:

$$(\boldsymbol{m}^\top\boldsymbol{\zeta})(\boldsymbol{m}^\top\mathbf{1}_m) = \begin{cases} 1, & \text{if } \exists j : \boldsymbol{m}_{[j]} = \boldsymbol{\zeta}_{[j]} = 1 \\ 0, & \text{otherwise} \end{cases}. \tag{29}$$

Note that (27) and (29) have the same formulation, thus the following equation holds:

$$(\boldsymbol{m}^\top\boldsymbol{\zeta})(\boldsymbol{m}^\top\mathbf{1}_m) = \boldsymbol{m}^\top\boldsymbol{\zeta} \tag{30}$$

$$\Rightarrow \quad (\mathbf{M}_{st}\boldsymbol{\zeta}) \odot (\mathbf{M}_{st}\mathbf{1}_n) = \mathbf{M}_{st}\boldsymbol{\zeta}. \tag{31}$$

Due to the equality of (25) and (26), the following relationship can be derived:

$$\mathbf{Z}\hat{\boldsymbol{p}}_t = \mathbf{M}_{st}\boldsymbol{\zeta} = \mathrm{diag}(\mathbf{M}_{st}\boldsymbol{\zeta})\mathbf{1}_n = \mathbf{Z}\mathbf{1}_n, \tag{32}$$

which shows that the second term in (8) is 0:

$$\mathbf{Z}(\mathbf{1}_n - \hat{\mathbf{p}}_t) = \mathbf{0}_n \tag{33}$$

$$\Rightarrow \quad (\mathbf{1}_n - \hat{\mathbf{p}}_t)^\top\mathbf{Z}\log(\mathbf{1}_n - \boldsymbol{p}_s) = 0. \tag{34}$$

12. Comparison with Other Losses

A theoretical overview about current counting losses is briefly introduced in the first part of Sec. 2: *Related Works*. Tab. 4 presents the empirical comparison, and P2R achieves better performance. Besides, the main paper provides a brief comparison between P2R and P2P in the 2nd paragraph of Sec. 6.1, given that both are designed for point-based counting models.

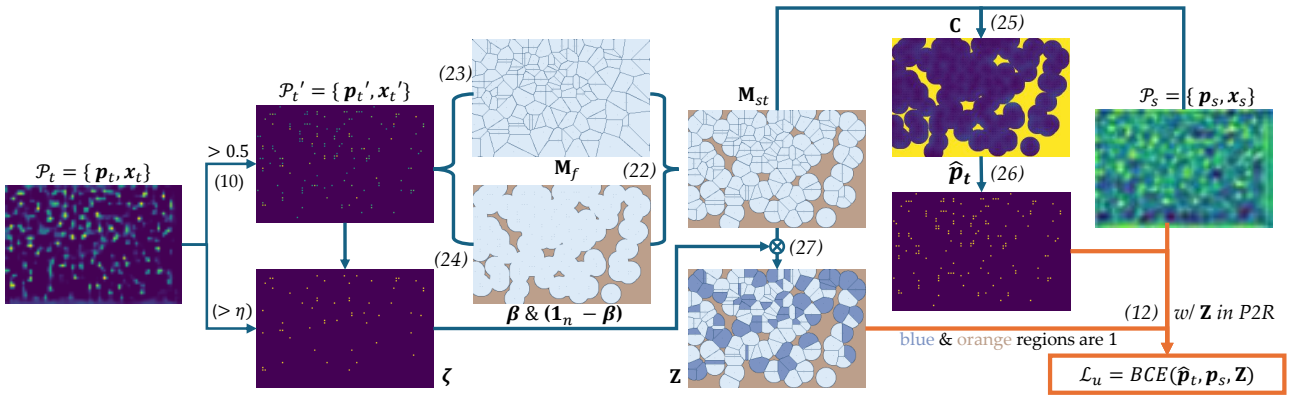
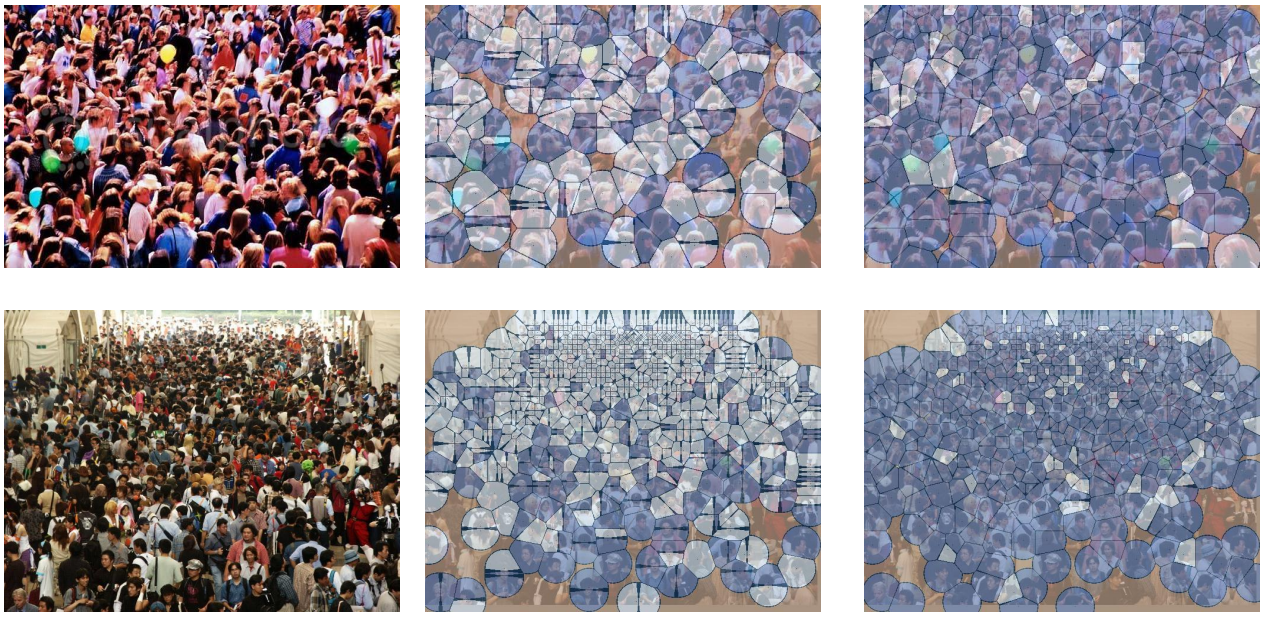


Figure 8. Computation of $\mathcal{L}_u$ in P2R.



(a) Input image (b) The 200th epoch (c) The last epoch

Figure 9. More visualization of the pseudo-labels.

13. Visualization

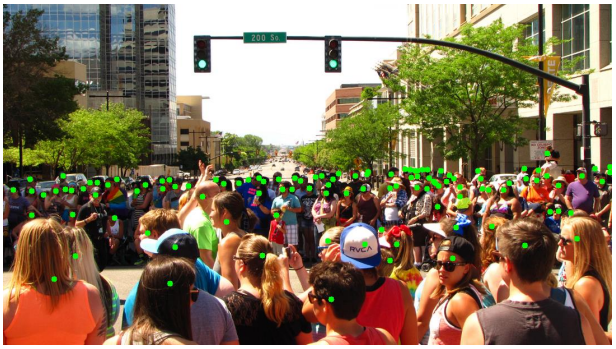
13.1. Pseudo-Labels

In Fig. 8, we illustrate the pipeline of loss computation for P2R. Given the teacher's prediction $\mathcal{P}_t$, the tensors $\mathcal{P}_t'$ and $\boldsymbol{\zeta}$ are generated by filtering out pixels with values greater than 0.5 and η, respectively. Subsequently, the region-to-point matching matrix $\mathbf{M}_{st}$ is constructed via (18), incorporating the results from the foreground assignment process (19) and the background definition (20). Consequently, the cost value to determine the learning objective within each region is estimated through (21). The learning objective $\hat{\boldsymbol{p}} = \hat{\mathbf{M}}\mathbf{1}$ is then defined, where $\hat{\mathbf{M}}$ identifies the potential foreground pixels in each region by (22). Finally, substituting $\boldsymbol{p}_s$, $\hat{\boldsymbol{p}}_t$, and the trustable region indicator $\mathbf{Z}$ obtained via (23) into the BCE loss (8) yields the final loss value for the P2R loss value in our semi-supervised crowd counting.

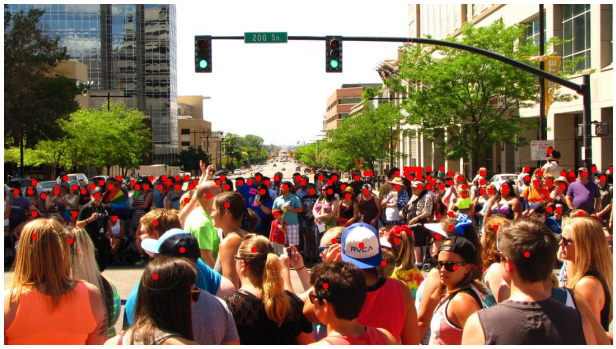
In the main paper, we visualize pseudo-labels in Figs. 4(e) and 4(f). In Fig. 9, we present two more examples to show the evolution of trusted regions from the 200th epoch to the end of training.

13.2. Comparison to GT

In Fig. 10, we present some visualization results of P2R when trained with 40% labeled data of the UCF-QNRF dataset. P2R can recognize the semantic information and localize pedestrians in the given images effectively.



GT count: 113



Pred count: 116

GT count: 2980

Pred count: 2832

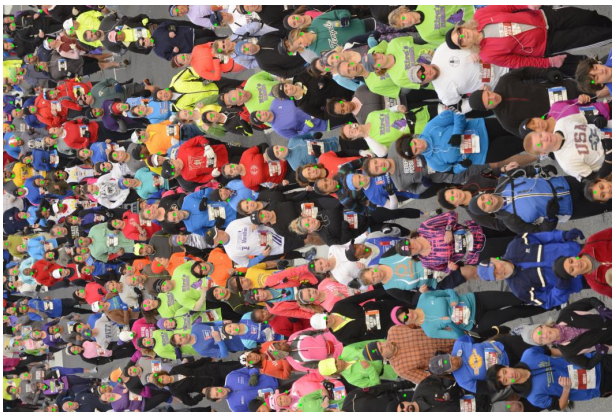
GT count: 150

Pred count: 143

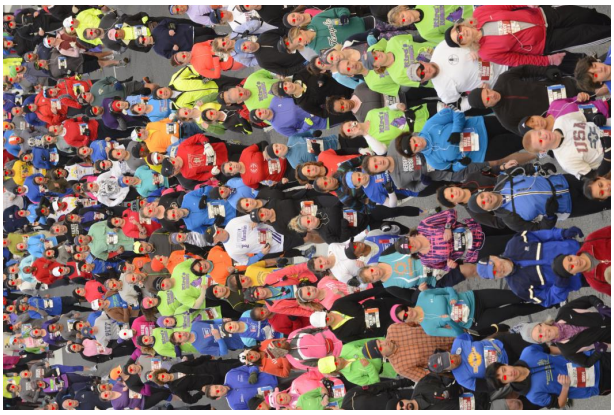
GT count: 1665

Pred count: 1602

Figure 10. Visualization of P2R's Prediction (# 1).



GT count: 172



Pred count: 168

GT count: 2441

Pred count: 2704

GT count: 312

Pred count: 318

GT count: 1550

Pred count: 1445

Figure 11. Visualization of P2R's Prediction (#2)