SCIENCE CHINA
Information Sciences

• LETTER •

April 2025, Vol. 68, Iss. 4, 149102:1–149102:2
https://doi.org/10.1007/s11432-024-4326-2

Self-calibrated region-level regression for crowd counting

Jiawen ZHU[1], Wenda ZHAO[1*], You HE[2] & Huchuan LU[1]

[1]*School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China*
[2]*Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*

Received 27 February 2024/Revised 5 October 2024/Accepted 8 January 2025/Published online 12 March 2025

Citation Zhu J W, Zhao W D, He Y, et al. Self-calibrated region-level regression for crowd counting. Sci China Inf Sci, 2025, 68(4): 149102, https://doi.org/10.1007/s11432-024-4326-2

Accurate crowd counting in natural images has become increasingly attractive owing to its numerous real-world applications, e.g., crowd analysis and video surveillance. Despite significant progress in crowd counting [1,2], challenges (such as scale variation and background clutter) remain.

To fully utilize spatial information, existing crowd counting approaches [3, 4] mainly estimate a density map, where point annotations are smoothed via a Gaussian kernel to generate probabilities indicating the presence of a crowd. The density map serves as the training objective, with pixel-level supervision applied, and the sum of the predicted density map is used for evaluation. However, strictly conducting pixel-level supervision based on the ground-truth density map has several drawbacks. (i) The size of the Gaussian kernel often fails to match crowd heads of varying sizes and occlusions, leading to the omission of valid content and the introduction of background clutter. (ii) Point annotations are not always marked at the exact center of heads, introducing additional errors in pixel-level supervision. (iii) The training objective focuses on pixel-to-pixel density maps, while the evaluation criterion is counts, creating an inconsistency [5]. To alleviate some of these issues, prior studies have proposed adaptive Gaussian kernels based on spatial distances to nearest neighbors, accommodating variations in crowd scale. However, these methods assume a uniform crowd distribution. Other approaches design dedicated loss functions to optimize pixel-level subregions with high disparities to the ground-truth density map, improving spatial awareness. Nevertheless, pixel-level density supervision remains a core component. Recently, some methods have introduced local counting maps to represent the crowd number within local patches, but these methods rely on fixed-size local counting regions.

To effectively leverage the spatial location information provided by point annotations, while avoiding the drawbacks of pixel-level density estimation, we propose a self-calibrated region (SCR) loss to calculate regional-level errors, achieving a proper balance between pixel-level and image-level counting. Considering the nonuniform distribution of crowd density, SCR loss dynamically determines and allocates subregion sizes based on the principle that sparse-crowd regions are larger, while dense-crowd regions are smaller. Furthermore, we introduce an unreliable margin attenuation strategy to alleviate the adverse effects in marginal regions.

Proposed method. We propose a self-calibrated layout method based on the principle that sparse-crowd regions are larger, while dense-crowd regions are smaller. The process is illustrated in Figure 1(a). Specifically, for a predicted density map, the two-dimensional plane can be divided both horizontally and vertically. First, we extract the crowd distribution along the x and y axes (see Figures 1(a)(ii) and (a)(iv)). To achieve this, we use matrices of ones, $\mathbf{1}_{H\times 1}$, $\mathbf{1}_{1\times W}$, with dimensions $H \times 1$ and $1 \times W$, to obtain the distribution of the number of people horizontally and vertically, respectively (see Figures 1(a)(iii) and (a)(v)):

$$\mathcal{S}_x = \mathcal{D}^{\mathrm{pr}} * \mathbf{1}_{H\times 1},\ \mathcal{S}_y = \mathbf{1}_{1\times W} * \mathcal{D}^{\mathrm{pr}}, \tag{1}$$

where W, H represent the width and height of the predicted density map $\mathcal{D}^{\mathrm{pr}}$, $\mathcal{S}_x \in \mathbb{R}^{W\times 1}$, $\mathcal{S}_y \in \mathbb{R}^{1\times H}$, and $*$ denotes the non-overlapping sliding convolution operation. Then, we decide on the layout according to the horizontal and vertical crowd density distributions. Here, we define two layout regulatory factors: γ_x and γ_y. We gain the dividing coordinates ζ_x and ζ_y in the x and y axes:

$$\zeta_x = \mathrm{Crop}(\mathcal{S}_x, \gamma_x H),\ \zeta_y = \mathrm{Crop}(\mathcal{S}_y, \gamma_y W), \tag{2}$$

where Crop refers to obtaining the x and y coordinates when the accumulated number of people along $\mathcal{S}_x$ and $\mathcal{S}_y$ reaches $\gamma_x H$ and $\gamma_y W$, respectively. For simplicity, we set γ_x equal to γ_y in this study, denoted as γ. It is worth mentioning that when $\gamma = 0$, our SCR loss is equivalent to the pixel-level loss. Conversely, when γ is sufficiently large, the size of the divided region becomes the entire image, and our SCR loss transitions to image-level regression loss. This design achieves a balance between pixel-level counting and image-level counting. Thus, the divided subregions of self-calibrated layout can be described as follows:

$$\{\mathcal{M}^{\mathrm{pr,sl}}_{i,k}\}_{k=1}^{K} = \mathrm{Div}(\mathcal{D}^{\mathrm{pr}}_i, \zeta_x, \zeta_y), \tag{3}$$

[*] Corresponding author (email: zhaowenda@dlut.edu.cn)

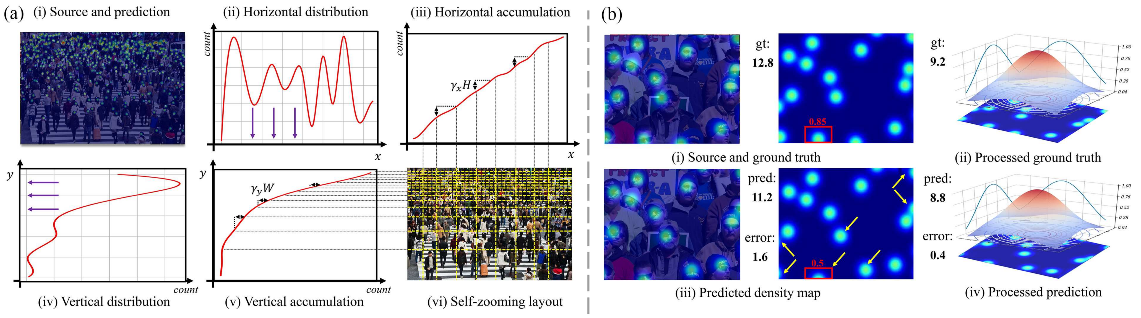
(a)
(i) Source and prediction
(ii) Horizontal distribution
count
x
(iii) Horizontal accumulation
count
$\gamma_x H$
x
(iv) Vertical distribution
y
count
(v) Vertical accumulation
y
$\gamma_y W$
count
(vi) Self-zooming layout
(b)
gt:
12.8
0.85
(i) Source and ground truth
gt:
9.2
(ii) Processed ground truth
pred:
11.2
error:
1.6
0.5
(iii) Predicted density map
pred:
8.8
error:
0.4
(iv) Processed prediction

Figure 1 (Color online) (a) Self-calibrated layout process; (b) illustration of the unreliable margin attenuation strategy.

$$\{\mathcal{M}_{i,k}^{\mathrm{gt,sl}}\}_{k=1}^{K} = \mathrm{Div}(\mathcal{D}_i^{\mathrm{gt}}, \zeta_x, \zeta_y). \tag{4}$$

The counting region is no longer set to a fixed pixel width and height but is instead divided adaptively based on the crowd distribution (as shown in Figure 1(a)(vi)). By contrast, the width and height of the image-level regression region are equal to those of the input image, while the width and height of the pixel-level regression region are set to 1. Clearly, after dividing adaptively, the size of the region in sparse-crowd areas will be larger, and vice versa for dense-crowd areas.

At the margins of the subregion grids, it is inevitable that a head may be divided across different regions. As a result, the ground-truth crowd number in these subregions becomes a discrete value. The ground-truth crowd numbers for these margin regions still depend on the ground-truth density map, which introduces drawbacks that can have adverse effects. For example, a region may visually appear to contain 0.5 heads, but the ground-truth counting value might be 0.85 (see Figures 1(b)(i) and (b)(iii)). While this discrepancy may seem minor for a single region, similar situations across all marginal regions can cumulatively result in significant errors. To address this issue, we propose an unreliable margin attenuation strategy to reduce this negative impact. The counting error in the k-th divided region is defined as $\Delta_k = \mathcal{C}_k^{\mathrm{pr}} - \mathcal{C}_k^{\mathrm{gt}}$, which is composed of errors from the center and the margin ($\Delta_k = \Delta_k^{\mathrm{cen}} + \Delta_k^{\mathrm{mar}}$). Based on this observation, the ground-truth crowd count at the center of a subregion is more reliable than that at the margins. In other words, the error calculated at the center (Δ_k^{cen}) is more trustworthy than that at the margin (Δ_k^{mar}). Therefore, we attenuate the unreliability error by applying a 2D Gaussian mask $\mathcal{G}(\cdot)$ centered at the location of each divided subregion to achieve the above processes (see Figure 1(b) for an illustration example). The weighted counting values in a divided region are formulated as follows:

$$\mathcal{C}_{i,k}^{\mathrm{pr,att}} = \sum \mathcal{M}_{i,k}^{\mathrm{pr,sl}} * \mathcal{G}(m, n; x_c, y_c; \sigma_w, \sigma_h), \tag{5}$$

$$\mathcal{C}_{i,k}^{\mathrm{gt,att}} = \sum \mathcal{M}_{i,k}^{\mathrm{gt,sl}} * \mathcal{G}(m, n; x_c, y_c; \sigma_w, \sigma_h), \tag{6}$$

$$\mathcal{G}(m, n; x, y; \sigma_w, \sigma_h) = \mathrm{e}^{-\left(\frac{(m-x)^2}{2\sigma_w^2} + \frac{(n-y)^2}{2\sigma_h^2}\right)}, \tag{7}$$

where x_c, y_c, w, h represent the center coordinates, width, and height of a divided region. Variances σ_w, σ_h of the Gaussian mask are proportional to the height and width of the divided region, where $\sigma_w = \lambda w + \epsilon$, $\sigma_h = \lambda h + \epsilon$, and λ and ϵ are set to 1.5 and 1.0 in all experiments. Through the above operation, we attenuate the weight of Δ_k^{mar} and obtain more reliable supervision. Thus, the final SCR loss is calculated as follows:

$$\mathcal{L}_{\mathrm{SCR}} = \frac{1}{2N} \sum_{i=1}^{N} \frac{1}{2K_i} \sum_{k=1}^{K_i} (\mathcal{C}_{i,k}^{\mathrm{pr,att}} - \mathcal{C}_{i,k}^{\mathrm{gt,att}})^2. \tag{8}$$

Experiments. We evaluate our proposed method by applying the SCR loss to multiple baseline methods while retaining other configurations. Comprehensive experimental results and analysis are provided in the supplementary material. Our method effectively improves counting performance, achieving competitive results compared to state-of-the-art approaches.

Conclusion. We propose a simple and straightforward region-level loss, termed SCR, for crowd counting. This loss dynamically divides the predicted density map into subregions based on the crowd density distribution, i.e., with larger layouts for sparse regions and smaller layouts for dense regions. Furthermore, we design an unreliable margin attenuation strategy to mitigate the influence of unreliable margins caused by region division, further improving the effectiveness of crowd counting. Extensive experiments conducted on five mainstream datasets demonstrate that our method significantly improves the performance of baseline methods and exhibits robustness to annotation inaccuracies. In the future, we aim to explore more flexible region-level learning strategies for crowd counting.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62176038, U1903215).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Zhang Y, Zhou D, Chen S, et al. Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 589–597
2 Li Y, Zhang X, Chen D. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1091–1100
3 Liu W, Salzmann M, Fua P. Context-aware crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 5099–5108
4 Du Z, Shi M, Deng J, et al. Redesigning multi-scale neural network for crowd counting. IEEE Trans Image Process, 2023, 32: 3664–3678
5 Liu X, Yang J, Ding W, et al. Adaptive mixture regression network with local counting map for crowd counting. In: Proceedings of the European Conference on Computer Vision, 2020. 241–257