Glance to Count: Learning to Rank with Anchors for Weakly-supervised Crowd Counting -Supplementary Materials-

Zheng Xiong[1,2], Liangyu Chai[1,2], Wenxi Liu[3], Yongtuo Liu[4], Sucheng Ren[2] and Shengfeng He[*2]

[1]South China University of Technology, [2]Singapore Management University
[3]Fuzhou University, [4]University of Amsterdam

In this supplement, we provide additional experiments and analyses for the proposed method.

1. Varied Degrees of Congestion

We further evaluate the proposed methods across varying levels of crowd congestion. Notably, we compare the performance of our method, referred to as *Ours*, with a baseline trained solely on regression labels. Our findings reveal that as crowd size increases (e.g., crowd number >1,000), *Ours* (Partial Weak Labels) and *Ours* exhibit improved performance. This highlights the adaptability of our model in dense scenes. We obtain similar outcomes on the UCF-QNRF dataset. Particularly in dense scenes, as illustrated in Fig 1, our proposed method *Ours* outperforms the baseline.

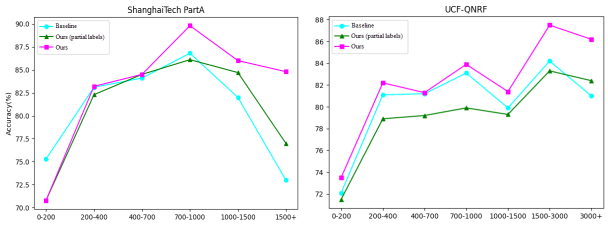

Figure 1: Accuracy (%) with different crowd numbers for the proposed models and baseline.

2. Beyond the "More than Twice" Assumption

We make a reasonable assumption that the number of people in the two images can be distinguished easily with more than twice the difference in crowd sizes between them. To explore more about the assumption, we incrementally adjust more critical times representing for difference in numbers between a pair of images by $1N, 1.5N, 2N, 3N$, where $2N$ represents the setting on the above assumptions in our experiments. The results are shown in Table 1. The results indicate that the ability of annotators to recognize and compare crowd numbers at a glance can affect the performance of the proposed method with a subjective aspect.

Table 1: Impacts of the crowd number difference between a ranking image pair. The best results are highlighted in bold.

Times	MAE	MSE
$1N$	**62.1**	**95.3**
$1.5N$	67.7	101.6
$2N$	69.8	104.2
$3N$	81.3	112.6

Table 2: Results under different crowd distributions.

	Uniformly		Non-Uniformly	
Method	**MAE**	**MSE**	**MAE**	**MSE**
Ours (Partial Labels w/o MLP)	66.4	105.3	72.5	118.4
Ours (Partial Labels w/ MLP)	66.2	104.8	70.6	113.1

3. Uniform and Non-uniform Distributions

To assess our performance across various scenarios, we conducted tests in settings featuring either uniform or non-uniform distributions. In the absence of a uniform dataset, we individually selected 50 crowd images from the ShanghaiTech Part A [6] dataset to construct datasets representing both uniform and non-uniform scenarios.

For uniform scenarios, we opted for crowd images showcasing an even distribution of people throughout the scene.

*Corresponding author. Email: shengfenghe@smu.edu.sg.

Table 3: Results of our proposed method in simple baseline(CSRNet [2]) on ShanghaiTech Part A [6] and UCF-QNRF [1]. "label level" refers to the supervision level of training. ✔ means the model employs all the labels under the corresponding level of supervision, and ✦ means the model employs a few labels at this supervision level. $*$ indicates the 0.1% of the parameters are tuned with location-level supervision. Note that, *Ours* exploits the same amount of count labels as other weakly supervised methods, and ranking labels can be auto-generated from count labels without extra annotating effort. ✯ indicates that our model is purely trained with ranking labels.

	Label level		**ST PartA**[6]		**UCF-QNRF**[1]	
Method	**Location**	**Count**	**MAE**↓	**MSE**↓	**MAE**↓	**MSE**↓
CSRNet [2]	✔		68.2	115.0	119.2	211.4
CSRNet (Count label only)		✔	85.6	128.1	149.0	245.3
CSRNet (Count label only+L2R[3])		✔	84.91	125.2	144.6	238.2
CCLS [5]		✔	104.6	145.2	-	-
Ours (Ranking Only)		✯	93.4	142.5	165.3	277.7
Ours (Partial Weak Labels)		✦	91.6	138.5	158.7	266.7
Ours		✔	76.9	113.1	133.0	218.8

For non-uniform scenarios, we chose images with more dispersed crowds and introduced random rotations to accentuate diverse distributions. The results of our method, Ours with Partial Labels, are presented in Table 2.

Our findings indicate that our model delivers superior performance on uniformly distributed data compared to non-uniform scenarios. This outcome aligns with the intuitive expectation that uniform distributions enable the algorithm to capture crowd features more evenly across the image. Moreover, we introduced the Upside-Down MLP branch to enhance the model's performance across different distributions. The results in Table 2 reveal that while the model with the MLP demonstrates slight improvement in uniform distribution scenarios, it significantly enhances performance in non-uniform scenarios. This underscores the module's effectiveness in enhancing the model's robustness across diverse distributions.

4. Effectiveness of Ranking

To assess the effectiveness of our ranking strategy, we utilize the CSRNet [2], a basic convolutional neural network (CNN)-based crowd counting architecture, as the backbone of the Siamese network. This backbone consists of a pretrained CNN for frontend 2D feature extraction and a dilated convolutional layer for expanding reception fields on the backend. For the sake of fairness, we omitted other proposed modules from the paper that could enhance performance.

As indicated in Table 3, CSRNet [2] denotes the baseline utilizing location-level supervision. CSRNet (Count label only) represents the regression-based baseline solely supervised by count labels. Adding to this, CSRNet (Count label only) +L2R [3] introduces "cropping rank" supervision to the previous label-only model. It is important to note that the weakly-supervised method CCLS [5], as well as our proposed method, both build upon the same feature extractor, CSRNet [2], for equitable comparison.

Evidently, our method outperforms the weakly-supervised regression method, CCLS [5], which represents the state-of-the-art approach employing regression-based training at the same supervision level. Moreover, our method, denoted as *Ours*(Ranking Only) and *Ours*(Partial Weak Labels), achieves lower counting error than the label-only baseline (CSRNet) due to the introduction of ranking pair supervision. Furthermore, it attains a lower error than the L2R supervised count label regression model. This demonstrates the considerable utility of ranking labels for regression in count-level supervision.

Importantly, *Ours* exhibits the best performance in a weakly-supervised setting in terms of MAE and MSE, approaching the performance of the location level baseline, CSRNet [2]. This underscores that even with a simple baseline, our method remains effective.

Table 4: The effect of ranking labels in the simulated real-world counting.

Setting	**MAE**	**MSE**
Baseline	101.8	138.2
Ours (Partial Weak Labels) (1,000 pairs)	83.3	124.5
Ours (Partial Weak Labels) (5,000 pairs)	79.4	119.1
Ours (Partial Weak Labels) (25,000 pairs)	77.4	112.3

5. Ranking Labels in Real-world Counting

Unlike the crowd data in benchmarks, the real-world crowd data is often a large quantity of unlabeled crowd images, so it is infeasible to annotate all of the extracted pairs. Especially in the mid-to-late stage of training, many abundant pairs with large numerical differences almost have no effect on loss, and just consume computing power in vain. Therefore, sparse labeling is preferable for massive, wild, and unlabeled crowd data. Denote the number of labels associated with crowd image x_i as $\zeta(x_i)$, and the average

number of $\zeta(x_i)$ in dataset D as $\zeta(D)$. (Note that only the labeled image will be included, thus $\zeta(\cdot) \geq 1$.) In short, with sufficient training samples, $\zeta(D)$ is closer to 1, which means the labels per image are sparse. Furthermore, an online labeling strategy can work well with sparse labeling. Before the start of training, the ranking pairs are labeled under the setting, $\zeta(D) = 1$. After training for some time, $\zeta(D)$ can become larger.

Practically, to verify the proposed ranking label in real scenes, we simulate the real annotation on a fixed number of real-world images. They are from JHU-CROWD++ [4], which contains 1.51 million annotated heads spanning 4,372 images, and is a challenging dataset with various scenarios. For the sparse labeling experiment, we randomly select 2,000 images forming 1,000 ranking pairs, and the sparse labeling strategy greatly reduces the number of labels per image $\zeta(D)$, from hundreds to one. In the other two experiments, we added annotated ranking pairs besides these 2,000 images with 5,000 pairs, and 25,000 pairs to verify the impact of label intensity.

The model is trained by combining ranking with free regression labels (*Ours* with partial labels) from the counting anchor set whose size is 50. The training is on the selected images from JHU-CROWD++, and the evaluation is on the widely used ShanghaiTech Part A dataset. As illustrated in Table 4, compared with the baseline of regression on the anchor set, the model with 1,000 comparison labels achieves a satisfactory performance with a few extra annotations. Adding more ranking labels leads to a slight improvement but brings a heavy annotation load.

References

[1] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *ECCV*, pages 532–546, 2018.

[2] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, pages 1091–1100, 2018.

[3] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *CVPR*, pages 7661–7669, 2018.

[4] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *ICCV*, pages 1221–1231, 2019.

[5] Yifan Yang, Guorong Li, Zhe Wu, Li Su, and Qingming Huang. Weakly-supervised crowd counting learns from sorting rather than locations. In *ECCV*. Springer International Publishing, 2020.

[6] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016.