Uniformity in Heterogeneity: Diving Deep into Count Interval Partition for Crowd Counting

Changan Wang[1*] Qingyu Song[1*] Boshen Zhang[1] Yabiao Wang[1]
Ying Tai[1] Xuyi Hu[1,3] Chengjie Wang[1†] Jilin Li[1] Jiayi Ma[4] Yang Wu[2]
[1]Tencent Youtu Lab, [2]Applied Research Center (ARC), Tencent PCG
[3]Department of Electronic & Electrical Engineering, University College London, United Kingdom
[4]Electronic Information School, Wuhan University, Wuhan, China
{changanwang, boshenzhang, caseywang, yingtai, jasoncjwang, jerolinli}@tencent.com,
qingyusong@zju.edu.cn, zceexhu@ucl.ac.uk, jyma2010@gmail.com, dylanywu@tencent.com

Supplementary

1. Inconsistent Ground-Truth Targets

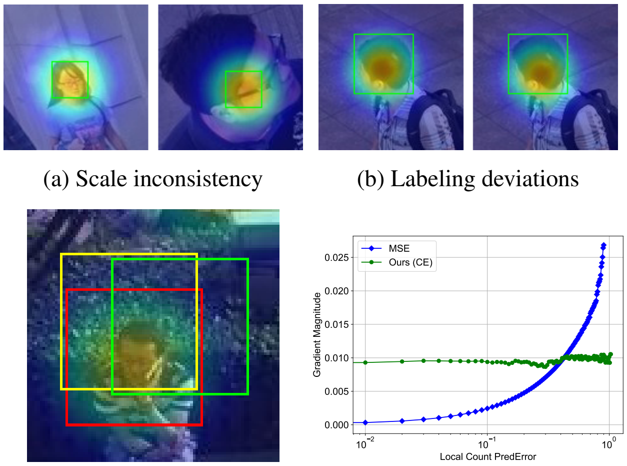
As shown in Figure 1, we list three types of inconsistencies between the semantic contents and the ground truth targets. These inconsistencies act as a kind of "noises" in the training targets, which might be harmful to the model learning.

2. Mathematical Analysis

Given an unseen testing image $\mathcal{I}$ without any prior, we calculate the expected counting error $\mathcal{E}$ for $\mathcal{I}$. Considering all possible K patches from the training set, there should be a collection $\mathcal{T}$ of local counts d_k, $k \in \{1, 2, ..., K\}$. We use $\widetilde{\mathcal{T}}$ to represent the collection after removing duplicate counts from $\mathcal{T}$. Assuming the data is independent and identically distributed (i.i.d.), then the local count map D_s of $\mathcal{I}$ can be viewed as another collection of local counts, which are randomly selected from $\widetilde{\mathcal{T}}$. Thus the error $\mathcal{E}$ for image $\mathcal{I}$ could be approximated as $\mathcal{E} \approx |\sum_{d_i \in \widetilde{\mathcal{T}}} p_i(d_i - \hat{d}_i)|$, in which p_i is the sampling probability for local count d_i, and $\hat{d}_i$ is the estimation for d_i. Typically, K is large enough so that p_i could be replaced with the frequency of occurrence N_{d_i}/K, and N_{d_i} is the number of occurrence for d_i in $\mathcal{T}$. Finally, the overall expected counting error for image $\mathcal{I}$ is represented as follows:

$$\mathcal{E} \approx \left| \sum_{d_i \in \widetilde{\mathcal{T}}} N_{d_i}(d_i - \hat{d}_i) \right| / K = \left| \sum_{k=1}^{K} (d_k - \hat{d}_k) \right| / K. \quad (1)$$

Since the expected counting error $\mathcal{E} \propto \tilde{\mathcal{E}} = |\sum_{k=1}^{K}(d_k - \hat{d}_k)|$, our goal is to minimize $\tilde{\mathcal{E}}$ with a suitable count interval partition and count proxy selection strategy. Assuming the ground truth count is $G = \sum_{k=1}^{K} d_k$, and the predicted count from the model is $\hat{G} = \sum_{k=1}^{K} \hat{d}_k$, in which $\hat{d}_k$ is the predicted count for local count d_k. $\hat{G}$ can be viewed as



(a) Scale inconsistency (b) Labeling deviations

(c) Semantic inconsistency (d) Robustness of Losses

Figure 1: Illustrations for three types of outliers introduced by the inconsistency between semantic content in patch and local count in ground-truth, and the comparison for robustness of MSE and CE. (a) Same local count but inconsistent semantic due to large scale variance, and the local counts in the two green boxes are the same but the latter one only covers parts of the head. (b) Same patch but different local counts due to labeling deviations. (c) Same head but different local counts for the three patches, which implies that different patches may have different local counts although they cover the same one head. (d) Compared with the robust CE loss, samples with larger prediction error contribute much larger gradients from the MSE loss, which might drown the useful and accurate gradients.

[*]Equal contribution. [†]Corresponding author.

two parts, $\hat{G}_{right}$ and $\hat{G}_{error}$. The former one is the predicted count when all d_k are classified correctly, and the latter one is the summation of the counting errors from all misclassified samples. Finally, the above goal of minimizing $\mathcal{E}$ should be converted to the problem of minimizing $\tilde{\mathcal{E}} = |G - (\hat{G}_{right} + \hat{G}_{error})|$.

The Mean Count Proxies Criterion. During testing stage, the count for a patch will be the proxy value δ_i if it is classified as the i-th interval c_i. Actually, when all the patches are classified correctly, *i.e.*, $\hat{G}_{error} = 0$, the $\tilde{\mathcal{E}}$ should represent the discretization errors due to the interval quantification. This can be demonstrated as follows:

$$\begin{aligned}\tilde{\mathcal{E}} &= |G - (\hat{G}_{right} + \hat{G}_{error})| = |G - \hat{G}_{right}| \\ &= |G - (n_1\delta_1 + n_2\delta_2 + ... + n_{m-1}\delta_{m-1})| \\ &= \left|\sum_{i=0}^{m-1}(x_{i1} + x_{i2} + ... + x_{in_i}) - \sum_{i=0}^{m-1} n_i\delta_i\right| \\ &= \left|\sum_{i=0}^{m-1}((x_{i1} + x_{i2} + ... + x_{in_i}) - n_i\delta_i)\right|.\end{aligned} \tag{2}$$

From the above equation, we could conclude that if we let $\delta_i = \sum_{j=1}^{n_i} x_{ij}/n_i$, $\tilde{\mathcal{E}}$ will get the minimal value 0. In other words, there will be *no extra quantization errors* when transforming the regression task into an interval classification problem, as long as we could choose a proper count proxy value δ_i for each interval. And the optimal count proxy is theoretically demonstrated as the average count value of samples in corresponding interval.

The Uniform Error Partition Criterion. According to the Equation 2, we have $|G - \hat{G}_{right}| = 0$ when using the proposed MCP criterion. Then we could derive that $\tilde{\mathcal{E}} = |G - (\hat{G}_{right} + \hat{G}_{error})| = |(G - \hat{G}_{right}) + \hat{G}_{error}| = |\hat{G}_{error}| = \left|\sum_{i=0}^{m-1} e_i\right|$, in which e_i is the counting error from the i-th interval due to misclassification. Obviously, it is nearly impossible to obtain a perfect model with all patches correctly classified. For a specific interval, the counting error depends on both the number of samples within the interval and the misclassification cost of each sample. Thus we try to minimize $\left|\sum_{i=0}^{m-1} e_i\right|$ with a comprehensive consideration of the above two factors.

We make further decomposition for e_i. Firstly, the misclassification counting error cost e_i is obviously proportional to the number of samples n_i. Secondly, for a single sample of interval c_i, it is more likely to be misclassified to a nearby interval c_j. And the corresponding error cost $e_{i\rightarrow j}$ is $\delta_j - \delta_i$, which is also approximately proportional to l_i since the interval lengths of adjacent intervals are nearly equal. In summary, $\tilde{\mathcal{E}} = \left|\sum_{i=0}^{m-1} e_i\right| \approx \alpha\left|\sum_{i=0}^{m-1} n_i l_i\right| \propto \left|\sum_{i=0}^{m-1} n_i l_i\right|$, in which we reasonably keep the constant α of all intervals the same for simplicity.

Intuitively, the UEP criterion makes the task of local count classification more easier to learn, yielding smaller prediction errors. Since the local count d_k in $\mathcal{T}$ follows a long-tailed distribution due to the extremely large density variation. If we only keep the same n_i for all intervals, the interval lengths of some intervals may be too large, which should lead to much larger misclassification error cost for them. Besides, if we keep the same l_i for all intervals, the sample number among intervals may be too unbalanced to train a well-performed classifier. Instead, the item $n_i l_i$ provides a good trade-off for the interval difficulty (*i.e.*, misclassification error cost) and the sample imbalance problem among intervals.

3. More Discussions

In this section, we conduct further discussions so that our approach can be better understood.

Further analysis on the effectiveness of IPH. From the ablation studies in the maintext, we find a relatively higher improvement for the IPH when using the multi-scale training. We provide a reasonable explanation as follows. With the augmentation of multi-scale training, relatively easier samples in the middle of each interval are optimized better, while the relatively harder samples around the interval borders become a performance bottleneck due to the ambiguity. On the contrary, after integrating with the IPH, the classification ambiguity for these harder samples is mitigated to some extent. In this way, these harder samples tend to benefit more from the multi-scale training, thus the performance bottleneck might be broken.

UEP is helpful for the prediction on background. Another key difference for count regression and our method is the way of dealing with background. Specifically, the paradigm of count regression learns an exact value 0 for the background. Such an approach has two disadvantages. Firstly, it cannot help the model to learn discriminative features, since all predictions less than 0 are equally considered as correct predictions due to the existence of ReLU activation before the output. Secondly, it is much more difficult to regress an accurate count, however a small regression error also matters due to large number of background samples. On the contrary, it is much easier to identify that if a background patch falls into the background interval in our method, thus avoiding the above problems. We further calculate the count error contribution ratios of the background for the two approaches under the same network structure. *The ratio is 10.21% for the MSE based regression model, and is only 1.73% for our model, which demonstrates the effectiveness of our method.*

Potential negative impacts of limited max local count. One may argue that the max count value is determined by the statistics in the training set, which might lead to poor generalization performance on unseen data. Let us clarify this issue from three aspects. Firstly, patches with extremely large local count are relatively rare due to the long-tailed distribution of local count. As a result, the counting errors from these patches should not contribute much to the final accuracy. Secondly, when the dataset is large enough, the training set and test set can be considered as Independent and Identically Distributed. In this circumstance, the max local count is equal for both training set and test set. Finally, the competitive results obviously clarify that the effectiveness of our method outweighs the negative impacts of limited max local count.

4. Visualized results

In this section, we present the visualized results of our method. Firstly, as shown in Table 1 and Table 2, our model performs very well under various crowd density. In particular, we observe an interesting phenomenon that our model seems to be able to better identify the fine-grained foreground regions compared with the ground-truth density map. This phenomenon implies that our model might have learned more discriminative information.

Secondly, we listed several cases where our model fails to accurately estimate the crowd number in Table 3. The regions with the worst prediction are marked with red rectangles. We group these cases into following three categories:

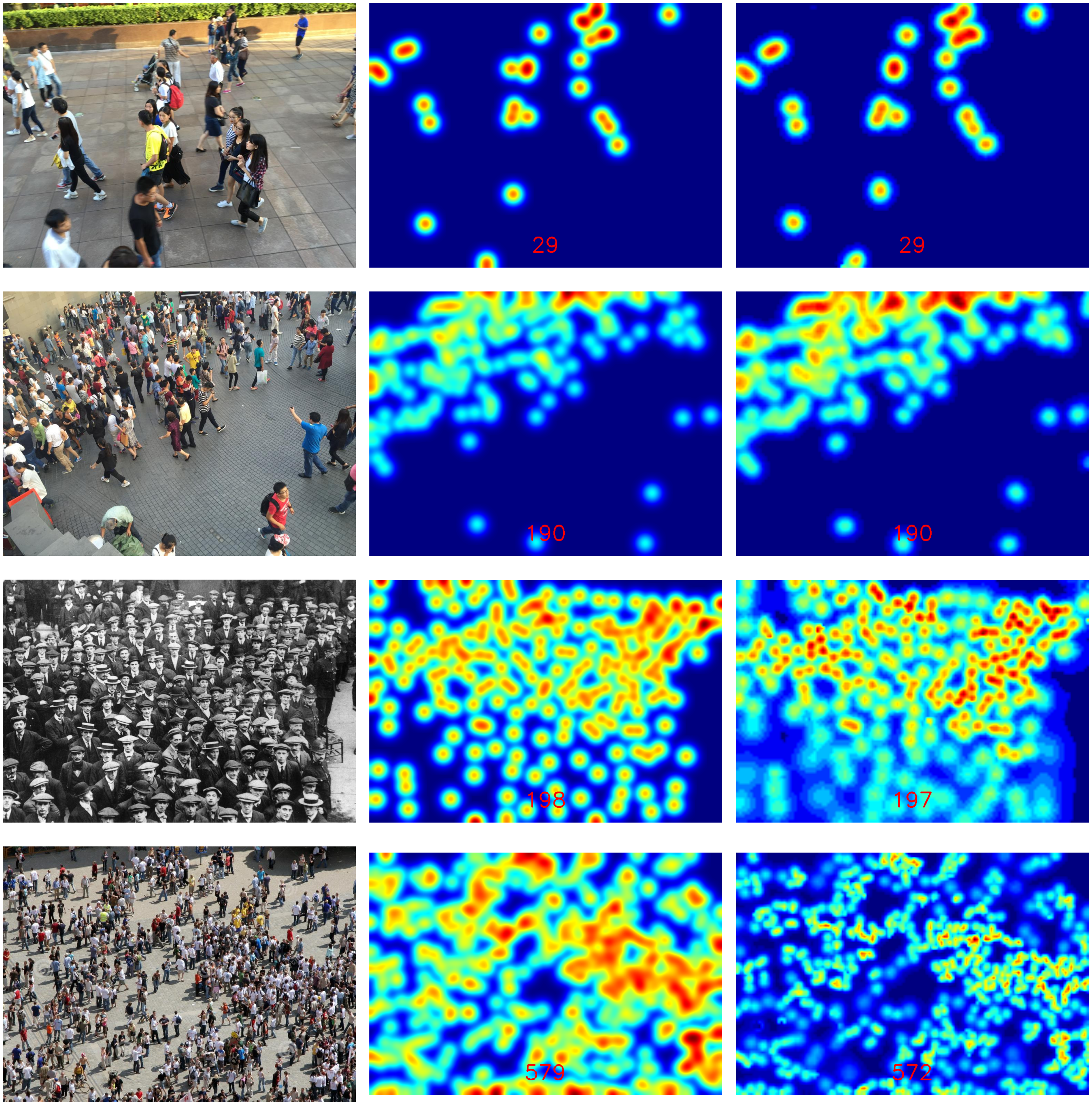
(1) Errors caused by missing annotations. As shown in the first row of Table 3, the missing annotation makes the ground-truth inaccurate. Strictly speaking, this should not be considered as a badcase, which however proves the superiority of our method in handling partial occlusions.

(2) Errors caused by severe occlusion. As shown in the second row of Table 3, the umbrella above the head makes it hard for our model to identify the boundary of the head.

(3) Errors caused by scarce training data. As shown in the third row and the fourth row of Table 3, insufficient data (night scenes and old photos) in the training set makes our model perform worse in such scenes.
Fortunately, all of the above errors could be alleviated to some extent by adding more training data.

5. Acknowledgments

We would like to thank Hao Lu for helpful discussions and suggestions.

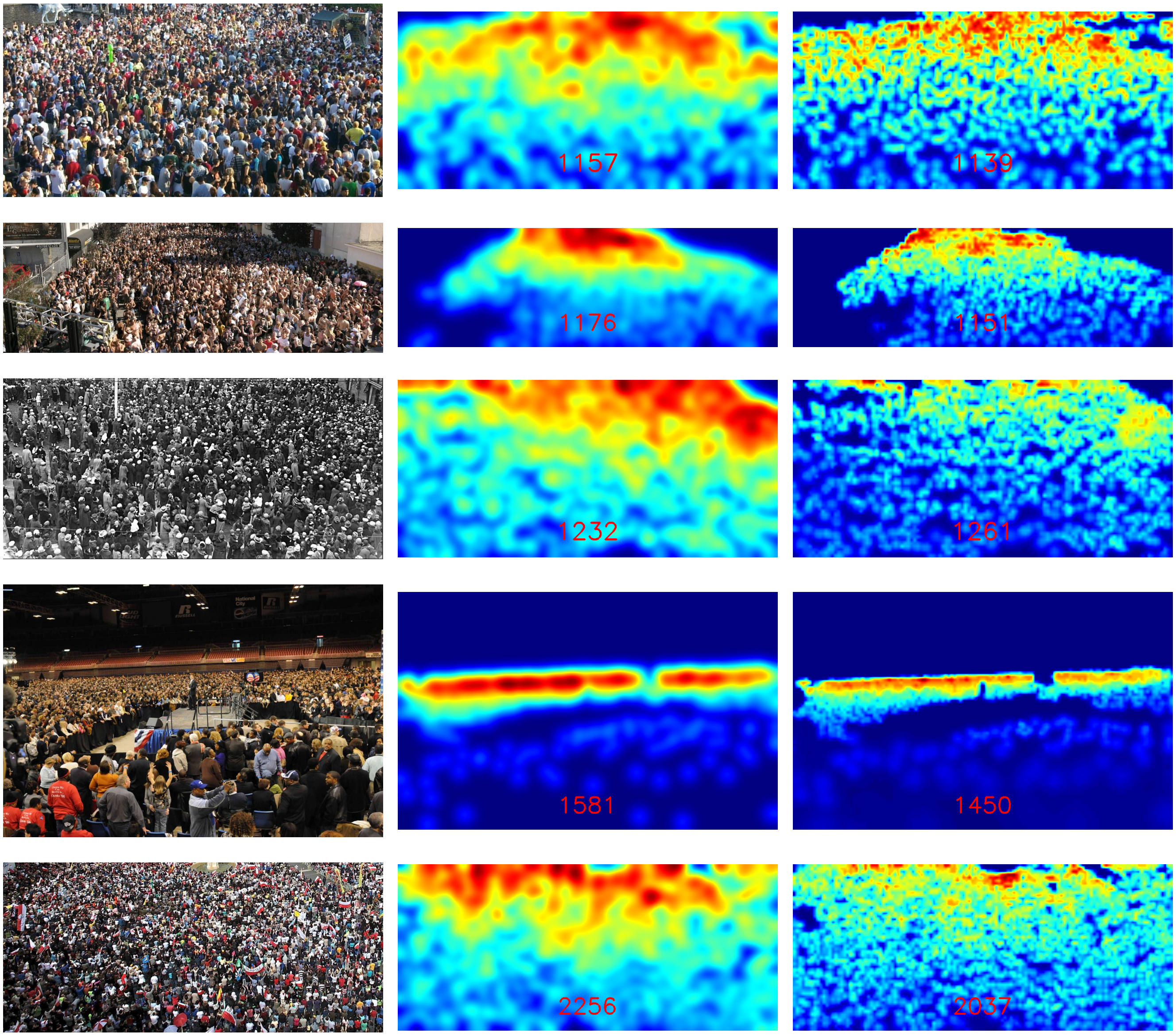


(a) Input Image

(b) Ground-Truth

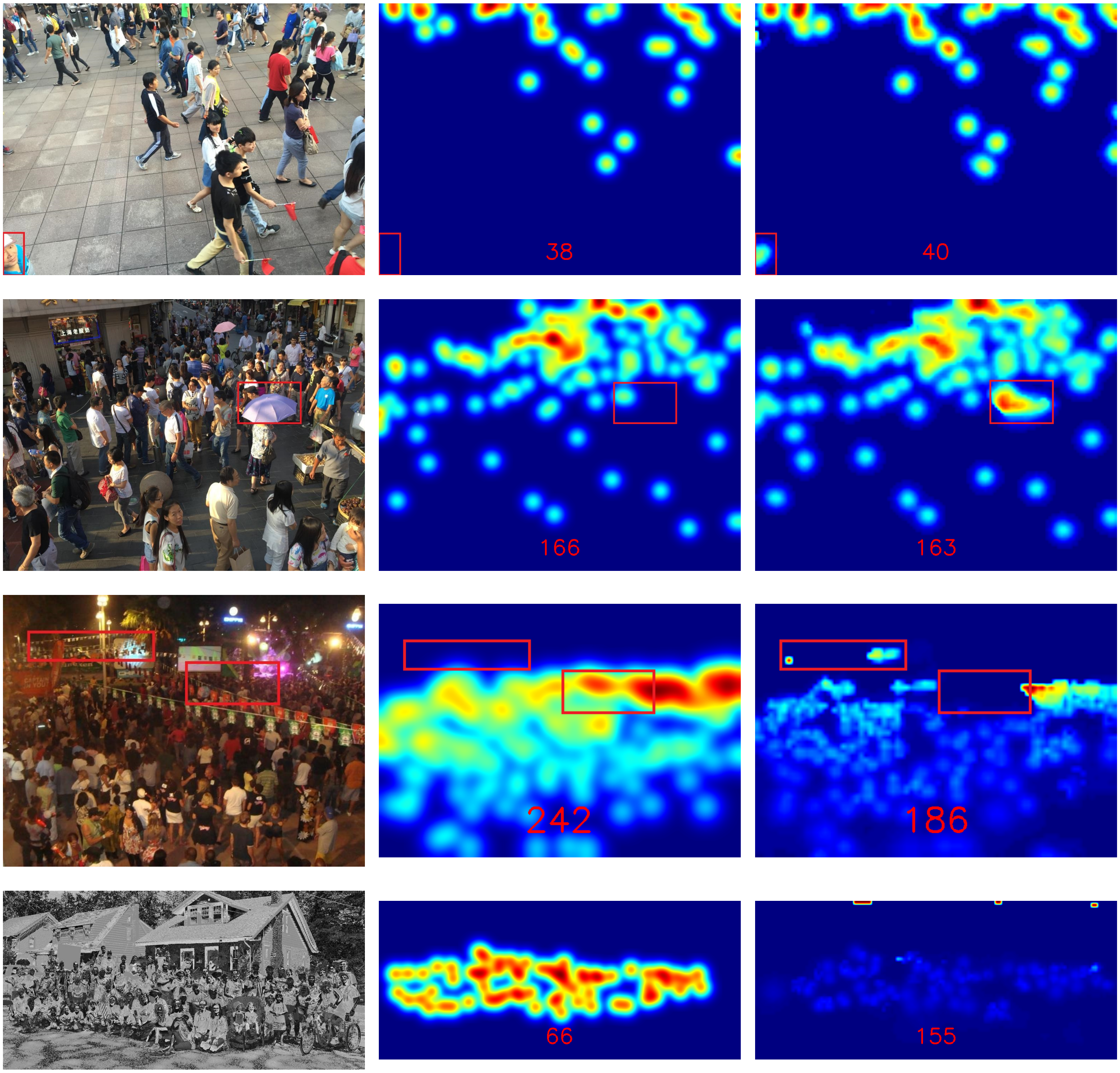
(c) Prediction of UEPNet

Table 1: Visualized results under sparse scenes.



(a) Input Image (b) Ground-Truth (c) Prediction of UEPNet

Table 2: Visualized results under congested scenes.



(a) Input Image

(b) Ground-Truth

(c) Prediction of UEPNet

Table 3: Visualized results for relatively bad cases. The regions with the worst prediction are marked with red rectangles.