

SEMI-SUPERVISED COUNTING VIA PIXEL-BY-PIXEL DENSITY DISTRIBUTION MODELLING

Anonymous authors

Paper under double-blind review

1 APPENDIX

In the appendix, we first study the influence of semi parameter (Sec. 1.1). Then we provide detailed alternative structures of dual-branch with tokens (Sec. 1.2). These alternatives and corresponding experimental results demonstrate the independence and effectiveness of P³Net structure. In Sec. 1.3 and Sec. 1.4, we study the counting accuracy of our method on the NWPU dataset and under the fully-supervised setting on five popular datasets respectively. Furthermore, we give a running cost evaluation (Sec. 1.5) and the pseudo code of P³Net learning process (Sec. 1.6). The visualizations of attention maps for each density token and a visual comparison with previous methods are shown in Sec 1.7. Finally, we introduce the setting of density interval partitions and label generation in Sec 1.8.

1.1 THE INFLUENCE OF PARAMETER λ

λ controls the proportion of gradient contribution of the unlabeled loss. As it gets larger, the self-supervised signals generated by unlabeled images will have greater effects on the model update. We hold experiments to study the influence of λ , which result is shown in Table 1.

The experiments are conducted on UCF-QNRF with a labeled ratio of 5%. As the result is shown, when λ is in the appropriate range, *i.e.* from 0.005 to 0.05, the gap of final counting performance is small. However, when the value is not suitable, the unlabeled images will impose too much or conversely, too little supervision on the gradient update of the model. Then the accuracy drops sharply, which is similar to excluding unlabeled data. Therefore, we set the unlabeled parameter $\lambda = 0.01$.

| λ | 0 | 0.0005 | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
|-----------|-------|--------|-------|-------|-------|-------|-------|
| MAE | 129.5 | 127.8 | 123.4 | 118.6 | 115.3 | 116.8 | 122.5 |
| MSE | 212.8 | 209.7 | 207.9 | 199.2 | 195.2 | 195.8 | 206.0 |

Table 1: The influence of parameter λ on UCF-QNRF (labeled ratio 5%).

1.2 DETAILED ALTERNATIVE STRUCTURE OF DUAL-BRANCH WITH TOKENS

To demonstrate the importance of the independence of model structures, we introduce four alternative structures. Specifically, they have the same CNN backbone, Pixel-wise Distribution Matching loss and inter-branch Expectation Consistency Regularization as P³Net. A general comparison of these structures is shown in Table 2. And the experimental results are shown in Table 3.

| | P ³ Net | SDDS | STDS | STSS | DTSS |
|--------------------------------|--------------------|------|------|------|------|
| Two Independent Decoders | ✓ | | ✓ | ✓ | ✓ |
| Two Series of Tokens | ✓ | ✓ | | | ✓ |
| Interleaving Density Semantics | ✓ | ✓ | ✓ | | |

Table 2: A comparison of structures for different alternatives.

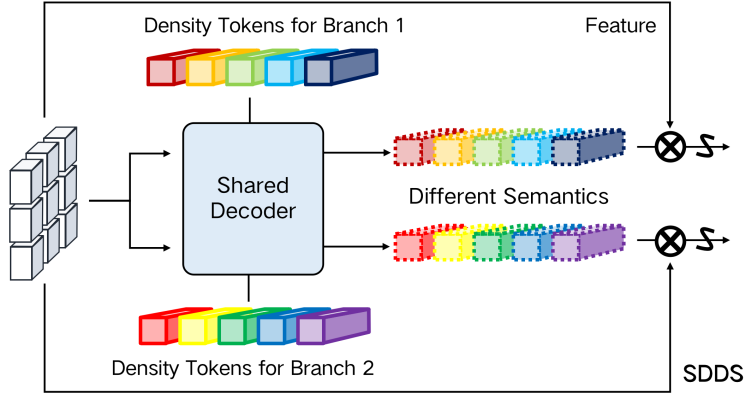


Figure 1: SDDS uses a shared decoder with two independent sets of tokens for interleaving intervals.

- ‘Shared Decoder with Different Semantics (SDDS)’ uses a common decoder to refine two interleaving and independent density tokens. Compared to P³Net, this alternative misses the independence of dual decoders for separate interaction between the feature map and the respective density tokens. The structure is shown in Figure 1. Since each token maintains exclusive semantic information, this structure still retains a certain degree of independence.
- ‘Shared Tokens with Different Semantics (STDS)’ uses independent dual decoders and a common series of density tokens with different semantics. The structure is shown in Figure 2. The alternative lacks the independence of different series of token features to represent interleaving density information. Instead, different semantics are endowed to the different decoders. Through the modulation and interaction of dual decoders, the refined tokens will fine-tune to be with interleaving density information. The drop in accuracy is attributed to the lack of semantic difference of initial density token prototypes in two interleaving branches.
- ‘Shared Tokens with Same Semantics (STSS)’ uses independent dual decoders but a common series of density tokens with same semantics. This structure, which is shown in Figure 3, lacks the independence of different series of representative token features and the interleaving density support. However, since the initialization of parameters in dual decoders is different, the confirmation bias will be partially eliminated, so the consistency regularization can still play a slight role.
- ‘Different Tokens with Same Semantics (DTSS)’ adopts the same independent model with P³Net but the dual density tokens represent the same density intervals. This alternative lacks the assumption that the semantics are exclusive and build on different supports, thus the consistency regularization is easier to satisfy. The structure is shown in Figure 4 and its semi-supervised performance deteriorates compared to that of P³Net.

| | SDDS | STDS | STSS | DTSS | P ³ Net |
|-----|-------|-------|-------|-------|--------------------|
| MAE | 123.3 | 121.4 | 123.7 | 123.1 | 115.3 |
| MSE | 215.7 | 197.8 | 213.4 | 206.0 | 195.2 |

Table 3: The impact of the settings of the consistency regularization with different structures. Experiments are conducted on UCF-QNRF with a labeled ratio 5%.

1.3 SEMI-SUPERVISED COUNTING PERFORMANCE ON NWPU

NWPU-CROWD Wang et al. (2020b) contains 5,109 images with 2.13 million annotated points. There are 3,109 images in training set, 500 images in validation set, and the remaining 1,500 images in testing set. The dataset has a large density range from 0 to 20,033 and contains various illumination scenes.

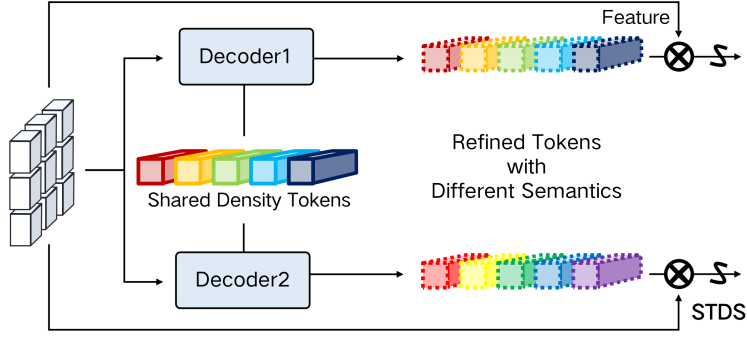


Figure 2: STDS uses two independent decoders with a shared set of tokens for interleaving intervals.

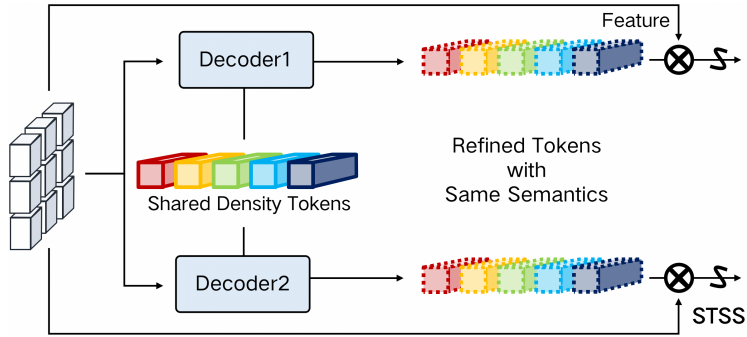


Figure 3: STSS uses two independent decoders with a shared set of tokens for the same intervals.

The semi-supervised setting on NWPU is referred to the work Meng et al. (2021). We keep the validation images to evaluate our model’s performance. In the training set, 10% images are randomly selected as the validation set. For the setting of labeled ratio of 5%, 10% and 40%, the corresponding proportion of images in the training set will be selected as labeled data and the rest images will be regarded as unlabeled data.

We compare our method with recent state-of-the-art semi-supervised methods, including mean teacher (MT) Tarvainen & Valpola (2017), Learning to Rank (L2R) Liu et al. (2018) and SUA Meng et al. (2021). The qualitative result is shown in Table 4. It can be observed that P³Net outperforms other methods by an obvious counting accuracy improvement on all three settings of ratios of labeled data.

| Labeled Ratio | 5% | | 10% | | 40% | |
|-------------------------------|--------------|--------------|-------------|--------------|-------------|--------------|
| | MAE | MSE | MAE | MSE | MAE | MSE |
| MT Tarvainen & Valpola (2017) | 184.0 | 648.0 | 144.1 | 508.6 | 129.8 | 515.0 |
| L2R Liu et al. (2018) | 159.2 | 650.3 | 138.3 | 550.2 | 125.0 | 501.9 |
| SUA Meng et al. (2021) | - | - | - | - | 111.7 | 443.2 |
| P ³ Net (Ours) | 116.7 | 598.8 | 88.2 | 515.9 | 76.3 | 422.8 |

Table 4: Comparisons with the state of the arts semi-supervised counting methods on NWPU. The experimental settings are referred to the work Meng et al. (2021).

1.4 PERFORMANCE UNDER FULLY-SUPERVISED SETTING

We hold experiments under the fully-supervised setting to study the effectiveness of proposed density tokens and Pixel-wise probabilistic Distribution (PDM) loss. It is worth mentioning that in

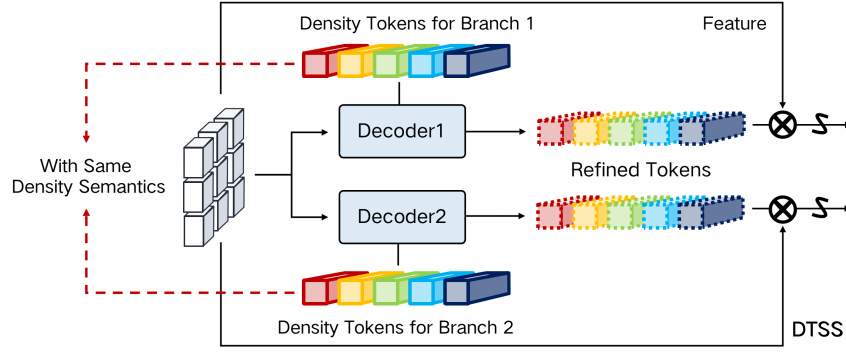


Figure 4: DTSS uses two independent decoders with two independent sets of tokens for the same intervals.

the fully-supervised setting, our model does not need the self-supervised inter-branch Expectation Consistency Regularization (ECR), which are design for learning from unlabeled data.

We compare our method with state-of-the-art fully-supervised methods on five datasets named UCF-QNRF, JHU-Crowd++, ShanghaiTech A, ShanghaiTech B and NWPU. The results are summarized in Table 5. Though it is not particularly designed for fully-supervised crowd counting, our method performs fairly well. For instance, on the UCF-QNRF dataset, it performs the best in terms of MAE and the second in terms of MSE, when compared with the most commonly used and best accepted methods like P2PNetSong et al. (2021), UEPNet Wang et al. (2021), DM-Count Wang et al. (2020a), BL Ma et al. (2019), etc. This consistent performance boost shows the effectiveness of the proposed architecture and the PDM loss. When considering the superior performance achieved by our method in semi-supervised crowd counting reported in the main test, it further implies that optimal semi-supervised counting is built on both the ability to learn from labeled data and unlabeled data.

| Dataset | UCF-QNRF | | JHU++ | | ShanghaiTech A | | ShanghaiTech B | | NWPU | |
|------------------------------|-------------|--------------|-------------|--------------|----------------|-------------|----------------|------------|-------------|--------------|
| Method | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE |
| MCNN Zhang et al. (2016) | 277 | 426 | 188.9 | 483.4 | 110.2 | 173.2 | 26.4 | 41.3 | 232.5 | 714.6 |
| CSRNet Li et al. (2018) | - | - | 85.9 | 309.2 | 68.2 | 115.0 | 10.6 | 16.0 | 121.3 | 387.8 |
| SANet Cao et al. (2018) | - | - | 91.1 | 320.4 | 67.0 | 104.5 | 8.4 | 13.6 | 190.6 | 491.4 |
| S-DCNet Xiong et al. (2019) | 104.4 | 176.1 | - | - | 58.3 | 95.0 | 6.7 | 10.7 | - | - |
| BL Ma et al. (2019) | 88.7 | 154.8 | 75.0 | 299.9 | 62.8 | 101.8 | 7.7 | 12.7 | 105.4 | 454.2 |
| DM-Count Wang et al. (2020a) | 85.6 | 148.3 | - | - | 59.7 | 95.7 | 7.4 | 11.8 | 88.4 | 388.6 |
| UOT Ma et al. (2021) | 83.3 | 142.3 | 60.5 | 252.7 | 58.1 | 95.9 | 6.5 | 10.2 | 87.8 | 387.5 |
| S3 Lin et al. (2021) | 80.6 | 139.8 | 59.4 | 244.0 | 57.0 | 96.0 | 6.3 | 10.6 | 83.5 | 346.9 |
| P2PNet Song et al. (2021) | 85.3 | 154.5 | - | - | 52.7 | 85.1 | 6.3 | 9.9 | 77.4 | 362.0 |
| UEPNet Wang et al. (2021) | 81.1 | 131.7 | - | - | 54.6 | 91.2 | 6.4 | 10.9 | - | - |
| Our method | 78.5 | 134.2 | 55.8 | 237.6 | 56.6 | 89.9 | 6.2 | 10.2 | 74.3 | 327.3 |

Table 5: Comparisons with the state of the arts on UCF-QNRF, JHU-Crowd++, ShanghaiTech A, ShanghaiTech B and NWPU under fully-supervised setting. The best performance is shown in **bold** and the second best is shown in underlined. Our model is very competitive with SOTA supervised methods.

1.5 RUNNING COST EVALUATION

We evaluate the running cost of our method, which comparison result is reported in Table 6. The result of floating-point operations (FLOPs) is computed on one 384×384 input image and the result of inference time is tested on 1024×1024 images. We compare P³Net with BL Ma et al. (2019) model which serves as a basic counting network, and VGG19+Trans where Trans stands for the vanilla transformer encoder with self-attention.

The growth of model parameters comes from the dual-branch decoder structure. However, without the need of regression head which is always composed of a convolutional network used in BL and VGG19+Trans, the FLOPs of P³Net achieve minimum. Meanwhile, with respect to feature length N and the token number $C \ll N$, the self attention in encoder takes $\mathcal{O}(N^2)$ time and space while the cross attention in decoder takes $\mathcal{O}(C^2 + CN) \approx \mathcal{O}(CN)$. From the results compared with VGG19+Trans, P³Net uses less FLOPs and inference time, justifying the low computational cost of decoder.

| | BL Ma et al. (2019) | VGG19+Trans | P ³ Net |
|---------------------------------|---------------------|-------------|--------------------|
| Model Size (M) | 21.5 | 29.9 | 36.8 |
| FLOPs (G) | 60.8 | 65.6 | 57.8 |
| Inference Time (s / 100 images) | 9.8 | 11.4 | 10.2 |

Table 6: Comparison of the Model Size, FLOPs and Inference Time. Trans stands for the vanilla encoder. The growth of model parameters comes from the dual-branch decoder structure. But without the need for a regression head which is always composed of a convolutional network, the number of floating-point operations executed (FLOPs) of P³Net is the minimum, and even lower than the the non-transformer-based BL Ma et al. (2019).

1.6 PSEUDO CODE

We provide a pseudo code for P³Net learning in Algorithm 1.

Algorithm 1: P³Net Learning

Input: Labeled dataset \mathcal{X} and unlabeled dataset \mathcal{U} .

Output: The counting model θ with dual density tokens $T = T_1, T_2$.

```

1 Initialize  $\mathcal{L} \leftarrow 0$ ;
2 for epoch in  $[1, maxepoch]$  do
3   for each sample  $s \in \mathcal{X} \cup \mathcal{U}$  do
4     Get the predicted distribution matrices  $O_1, O_2$  by corresponding  $T_1, T_2$ ;
5     if  $s \in \mathcal{X}$  then
6       Generate the training labels  $Y_1, Y_2$  by ground-truth;
7        $\mathcal{L} \leftarrow \mathcal{L}_P$  by calculating the PDM loss based on Eq. 7;
8     else
9       Generate the pixel-wise mask  $\mathcal{E}$  based on Eq. 9;
10       $\mathcal{L} \leftarrow \lambda \mathcal{L}_E$  by calculating the ECR loss based on Eq. 8;
11    end
12    Update the counting model  $R$  and density tokens  $T$  minimizing  $L$ ;
13  end
14 end
15 Return the trained counting model  $\theta'$  with  $T'$ .
```

1.7 VISUALIZATIONS

We visualize the attention maps for each density token in dual branch to study their effects. Visualizations are shown in Figure 5 for a labeled image and Figure 6 for an unlabeled image.

The attention maps are generated by the same model, which is trained on UCF-QNRF with a labeled ratio of 5%. A and B stand for different branch and the numbers represent the different tokens. Tokens with higher numbers specify the density interval with higher density. The quantitative results show that the density tokens work well whether the predicted density was supervised by a clear ground-truth label or not.

We also provide visual comparisons between our model and the previous state-of-the-art semi-supervised counting model SUA (Meng et al., 2021) in Figure 7. We visualize the predicted densities

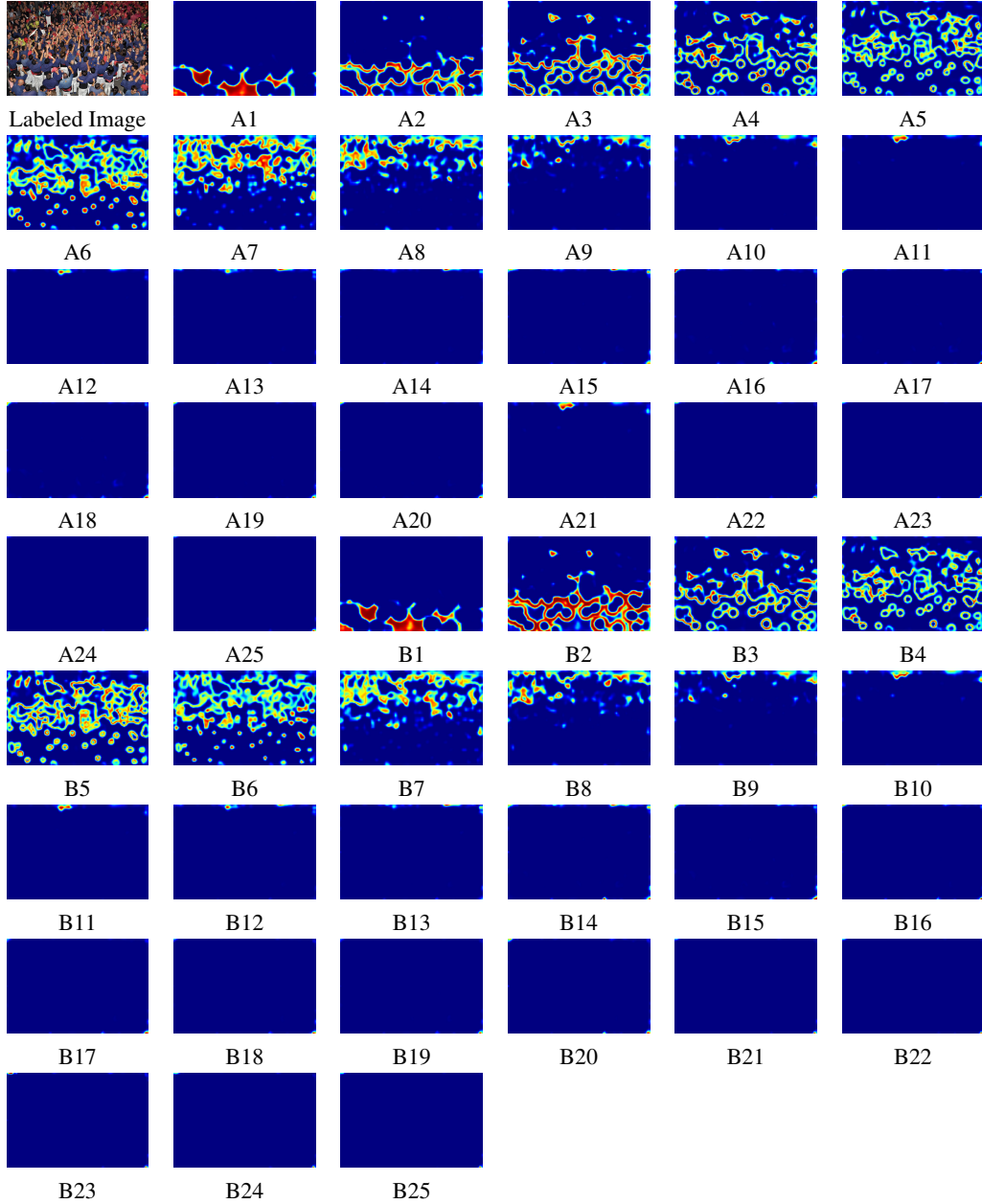


Figure 5: Attention maps of each density token in dual branch for a labeled training image when training with a labeled ratio of 5% on UCF-QNRF. A and B stand for different branch and the numbers represent the different tokens. Tokens with higher numbers specify the density interval with higher density.

on unlabeled training images of ShanghaiTech A of both models. The first row presents input images. The second row presents predicted density maps by SUA model while the third row presents predicted density maps by our P³Net. For SUA model in unlabeled data, there will be serious false alarms in the background. In contrast, our density token guided model can perform more stable and thus produce density maps with better accuracy.

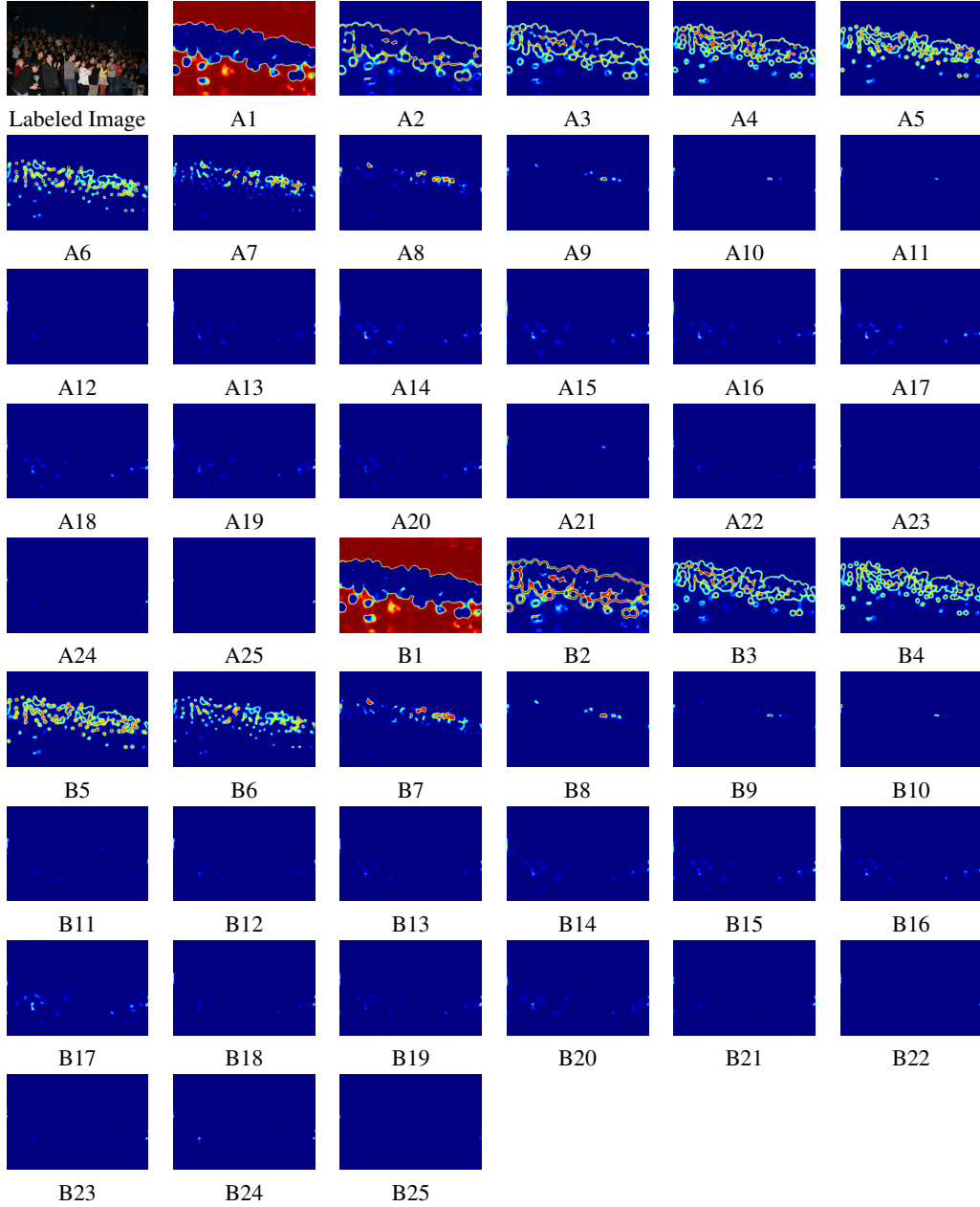


Figure 6: Attention maps of each density token in dual branch for an unlabeled training image when training with a labeled ratio of 5% on UCF-QNRF. A and B stand for different branch and the numbers represent the different tokens. A token with a higher number specifies the density interval with higher density.

1.8 THE SETTING OF DENSITY INTERVAL PARTITIONS AND LABEL GENERATION

The interval partitions are preset and remain constant during training and inference stages. We follow the paper (Wang et al., 2021) to obtain the appropriate intervals.

The partitions for the first branch are

[0, 0.0019, 0.0081, 0.0165, 0.0272, 0.0404, 0.056, 0.076, 0.099, 0.126, 0.159, 0.199, 0.246, 0.303, 0.371, 0.454, 0.556, 0.684, 0.848, 1.06, 1.36, 1.8, 2.5, 3.9, 8.2].

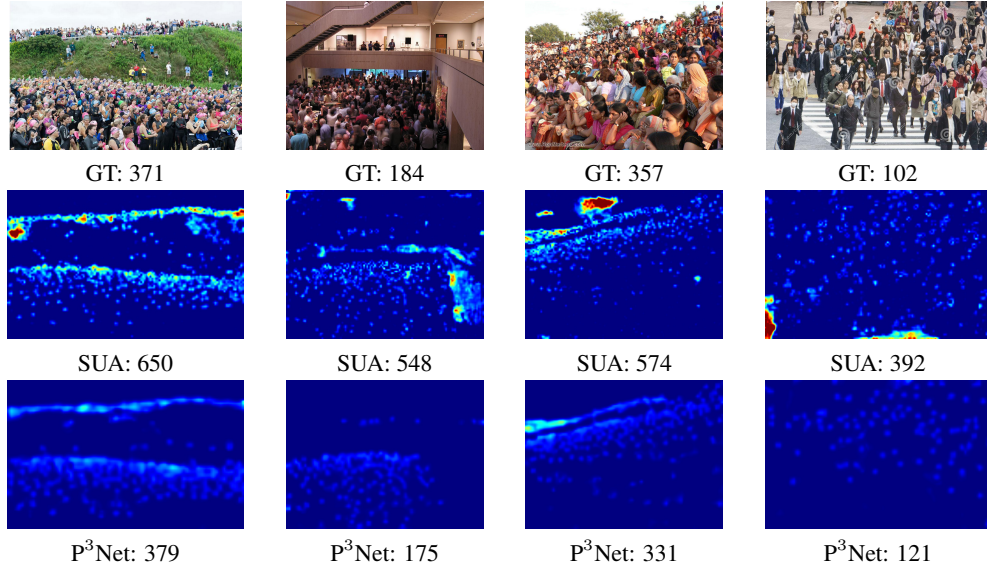


Figure 7: Visualizations of predicted densities on unlabeled training images of ShanghaiTech A. The first row: input images. The second row: predicted density maps by SUA model. The third row: predicted density maps by our P³Net. For SUA model in unlabeled data, serious false alarms in the background are observed, as shown in the second row. In contrast, our density token guided model can perform more stable and thus produce density maps with better accuracy in the third row.

116 The partitions for the second branch are

117 [0, 0.00087, 0.0046, 0.0119, 0.0214, 0.0333, 0.048, 0.065, 0.086, 0.112, 0.142, 0.178, 0.221, 0.272,
118 0.334, 0.409, 0.501, 0.615, 0.759, 0.945, 1.197, 1.55, 2.1, 3.0, 4.5, 8.5].

119 As the original ground-truth for the counting task is in the form of discrete points. To generate labels
120 for PDM loss, we first take the most popular density map generation solutions as the paper (Zhang
121 et al., 2016), smoothes each ground-truth point by a 2D Gaussian kernel. Then we calculate the
122 total density in each patch and assign them into corresponding intervals, which are pre-defined by
123 discretizing the whole density space.

| | |
|--|-------------------|
| REFERENCES | 124 |
| Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In <i>ECCV</i> , 2018. | 125 126 |
| Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In <i>CVPR</i> , 2018. | 127 128 |
| Hui Lin, Xiaopeng Hong, Zhiheng Ma, Xing Wei, Yunfeng Qiu, Yaowei Wang, and Yihong Gong. Direct measure matching for crowd counting. <i>IJCAI</i> , 2021. | 129 130 |
| Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In <i>CVPR</i> , 2018. | 131 132 |
| Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In <i>ICCV</i> , 2019. | 133 134 |
| Zhiheng Ma, Xing Wei, Xiaopeng Hong, Hui Lin, Yunfeng Qiu, and Yihong Gong. Learning to count via unbalanced optimal transport. In <i>AAAI</i> , 2021. | 135 136 |
| Yanda Meng, Hongrun Zhang, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Huang, and Yalin Zheng. Spatial uncertainty-aware semi-supervised crowd counting. In <i>ICCV</i> , 2021. | 137 138 |
| Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In <i>ICCV</i> , 2021. | 139 140 141 |
| Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. <i>NIPS</i> , 2017. | 142 143 |
| Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. <i>NIPS</i> , 2020a. | 144 145 |
| Changan Wang, Qingyu Song, Boshen Zhang, Yabiao Wang, Ying Tai, Xuyi Hu, Chengjie Wang, Jilin Li, Jiayi Ma, and Yang Wu. Uniformity in heterogeneity: Diving deep into count interval partition for crowd counting. In <i>ICCV</i> , 2021. | 146 147 148 |
| Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. <i>PAMI</i> , 2020b. | 149 150 |
| Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In <i>ICCV</i> , 2019. | 151 152 |
| Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In <i>CVPR</i> , 2016. | 153 154 |