

From Open Set to Closed Set: Supervised Spatial Divide-and-Conquer for Object Counting

Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Chunhua Shen, Zhiguo Cao

Abstract—Visual counting, a task that aims to estimate the number of objects from an image/video, is an open-set problem by nature as the number of population can vary in $[0, +\infty)$ in theory. However, collected data are limited in reality, which means that only a closed set is observed. Existing methods typically model this task through regression, while they are prone to suffer from unseen scenes with counts out of the scope of the closed set. In fact, counting has an interesting and exclusive property—spatially decomposable. A dense region can always be divided until sub-region counts are within the previously observed closed set. We therefore introduce the idea of spatial divide-and-conquer (S-DC) that transforms open-set counting into a closed set problem. This idea is implemented by a novel Supervised Spatial Divide-and-Conquer Network (SS-DCNet). It can learn from a closed set but generalize to open-set scenarios via S-DC. We provide theoretical analyses and a controlled experiment on synthetic data, demonstrating why closed-set modeling works well. Experiments show that SS-DCNet achieves state-of-the-art performance in crowd counting, vehicle counting and plant counting. SS-DCNet also demonstrates superior transferability under the cross-dataset setting. Code and models are available at: <https://git.io/SS-DCNet>.

Index Terms—Object Counting, Open Set, Closed Set, Spatial Divide-and-Conquer

1 INTRODUCTION

Counting is an open-set problem by nature as a count value can range from 0 to $+\infty$ in theory. It is therefore typically modeled in a regression manner. Benefiting from the success of convolutional neural networks (CNNs), state-of-the-art deep counting networks often adopt a multi-branch architecture to enhance the feature robustness to dense regions in an image [2], [3], [51]. However, the observed patterns in datasets are limited in practice, which means that networks can only learn from a *closed set*. Are these counting networks still able to produce accurate predictions when *the number of objects is out of the scope of the closed set*? According to Fig. 2, local counts observed in the closed set exhibit a long-tailed distribution. Extremely dense patches are rare while sparse patches take up the majority. As what can be observed, increased local density leads to significantly deteriorated performance in relative mean absolute error (rMAE). *Is it necessary to set the working range of CNN-based counters to the maximum count value observed, even though a majority of samples are sparse and the counter works poorly in this range?*

In fact, counting has an interesting and exclusive property—being spatially decomposable. The above problem can be largely alleviated with the idea of spatial divide-and-conquer (S-DC). Suppose that a network has been trained

This work is supported by the Natural Science Foundation of China under Grant No. 61876211. (Corresponding author: Zhiguo Cao. H. Xiong and H. Lu contributed equally. Part of the work was done when H. Xiong was visiting The University of Adelaide.)

H. Xiong, C. Liu, L. Liu, Z. Cao are with the National Key Laboratory of Science and Technology on Multi-Spectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: hpxiong@hust.edu.cn, cx_liu@hust.edu.cn, wings@hust.edu.cn, zgcao@hust.edu.cn).

H. Lu and C. Shen are with the University of Adelaide, SA 5005, Australia (e-mail: hao.lu@adelaide.edu.au, chunhua.shen@adelaide.edu.au).



Figure 1. An illustration of spatial divisions. Suppose that the closed set of counts is $[0, 20]$. In this example, dividing the image for one time is inadequate to ensure that all sub-region counts are within the closed set. For the top left sub-region, it needs a further division.

to accurately predict a closed set of counts, say $0 \sim 20$. When facing an image with extremely dense objects, one can keep dividing the image into sub-images until all sub-region counts are less than 20. The network can then count these sub-images and sum over all local counts to obtain the global image count. Fig. 1 depicts the overall idea of S-DC. A subsequent question is how to spatially divide the count. A naive approach is to upsample the input image, divide it into sub-images and process sub-images with the same network. This approach, however, is likely to blur the image and lead to exponentially-increased computation cost and memory consumption. Inspired by fully convolutional networks and ROI pooling [12], we show that it is feasible to achieve S-DC on feature maps, as shown in Fig. 3. By decoding and upsampling the feature map, the following prediction layers can focus on the feature of local areas and predict sub-region counts accordingly.

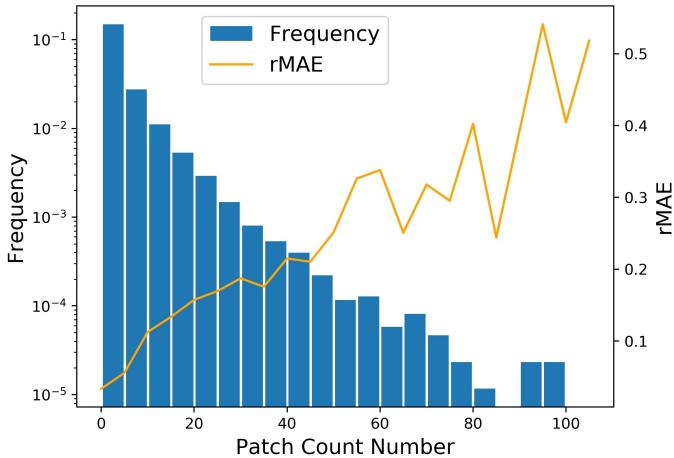


Figure 2. The histogram of count values of 64×64 local patches on the test set of ShanghaiTech Part_A dataset [51]. The orange curve denotes the relative mean absolute error (rMAE) of CSRNet [21] on local patches.

To implement the idea above, we propose a simple yet effective Supervised Spatial Divide-and-Conquer Network (SS-DCNet). SS-DCNet learns from a closed set of count values but is able to generalize to open-set scenarios. Specifically, SS-DCNet adopts a VGG16 [38]-based encoder and an UNet [34]-like decoder to generate multi-resolution feature maps. All feature maps share the same counter. The counter can be designed by following the standard local count regression paradigm [25] or by discretizing continuous count values into a set of intervals as a classifier following [19], [23]. Furthermore, a division decision module is designed to decide which sub-region should be divided and to merge different levels of sub-region counts into the global image count.

We provide theoretical analyses to shed light on why the transition from the open set to the closed set makes sense for counting. We also show through a controlled experiment on synthetic data that, even given a closed training set, SS-DCNet effectively generalizes to the open test set. The effectiveness of SS-DCNet is further demonstrated on three crowd counting datasets (ShanghaiTech [51], UCF_CC_50 [14] and UCF-QNRF [15]), a vehicle counting dataset (TRANCOS [13]), and a plant counting dataset (MTC [25]). Results show that SS-DCNet indicates a clear advantage over other competitors and sets the new state of the art. In addition, we remark that the closed set of SS-DCNet executes an implicit transfer in the output space, which is backed by state-of-the-art performance under the cross-domain evaluations. In particular, SS-DCNet even beats most state-of-the-art counting models that are trained directly on the target domain in the task from UCF-QNRF to ShanghaiTech Part_A.

In summary, the main contributions of this work are as follows.

- We propose to transform open-set counting into a closed-set problem via S-DC. A theoretical analysis of why such a transformation works well is also presented;
- We investigate the explicit supervision for S-DC, which leads to a novel SS-DCNet. SS-DCNet is applicable to both regression-based and classification-based counters and can produce visually clear spatial divisions;

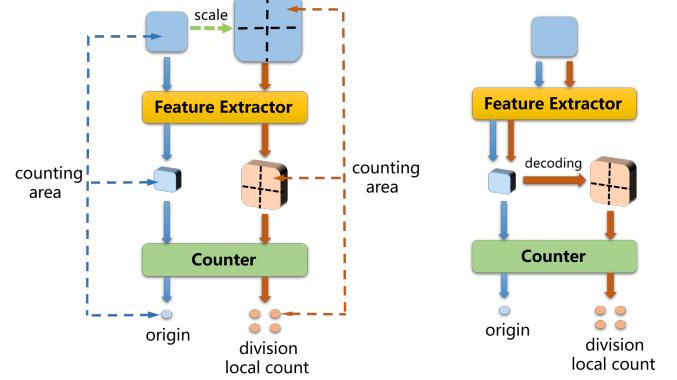


Figure 3. Spatial divisions on the input image (left) and the feature map (right). Spatially dividing the input image is straightforward. The image is upsampled and fed to the same network to infer counts of local areas. The orange dashed line is used to connect the local feature map, the local count and the sub-image. S-DC on the feature map avoids redundant computations and is achieved by upsampling, decoding and dividing the feature map of high resolution.

- We report state-of-the-art counting performance over 5 challenging datasets with remarkable relative improvements. We also show good transferability of SS-DCNet via cross-dataset evaluations on crowd counting datasets.

A preliminary conference version of this work appeared in [47] where S-DCNet, the first version of SS-DCNet, was developed. Here we have extended [47] in the following aspects: i) we provide theoretical analyses why closed set modeling works well; ii) we further enhance S-DCNet at the methodology level by investigating further a regression-based closed-set counter, by integrating a count-orientated upsampling operator and by improving the model training with explicit supervision of spatial divisions; iii) we provide more ablative studies and qualitative analyses to highlight the role of S-DC; and iv) we give an insight of SS-DCNet w.r.t. its good transferability in the output space and report state-of-the-art performance under the cross-dataset evaluation setup.

2 RELATED WORK

Current CNN-based counting approaches are mainly built upon the framework of local regression. According to their regression targets, they can be categorized into two categories: density map regression and local count regression. We first review these two regression paradigms. Since SS-DCNet works not only in regression counts but also in classification, some works that reformulate the regression problem are also discussed.

2.1 Density Map Regression

The concept of density map was introduced in [18]. The density map contains the spatial distribution of objects, thus can be smoothly regressed. Zhang *et al.* [50] may be the first to adopt a CNN to regress local density maps. Then almost all subsequent counting networks followed this idea. Among them, a typical network architecture is multi-branch. MCNN [51] and Switching-CNN [2] used three columns of CNNs with varying receptive fields to depict objects

of different scales. SANet [3] adopted Inception [42]-liked modules to integrate extra branches. CP-CNN [39] added two extra density-level prediction branches to combine global and local contextual information. ACSCP [36] inserted a child branch to match cross-scale consistency and an adversarial branch to attenuate the blurring effect of the density map. ic-CNN [33] incorporated two branches to generate high-quality density maps in a coarse-to-fine manner. IG-CNN [1] and D-ConvNet [37] drew inspirations from ensemble learning and trained a series of networks or regressors to tackle different scenes. DecideNet [22] attempted to selectively fuse the results of density map estimation and object detection for different scenes. Unlike multi-branch approaches, Idrees *et al.* [15] employed a composition loss and simultaneously solved several counting-related tasks to assist counting. CSRNet [21] benefited from dilated convolution which effectively expanded the receptive field to capture contextual information.

Existing deep counting networks aim to generate high-quality density maps. However, density maps are actually in the open set as well. For a single point, different kernel sizes lead to different density values. When multiple objects exist and are close, density patterns are even much diverse. Since observed samples are limited, density maps are clearly in an open set. In addition, density maps do not have the physical property of spatial decomposition. We therefore cannot apply S-DC to density maps.

2.2 Local Count Regression

Local count regression directly predicts count values of local image patches. This idea first appeared in [6] where a multi-output regression model was used to regress region-wise local counts simultaneously. Authors of [9] and [25] introduced such an idea into deep counting. Local patches were first densely sampled in a sliding-window manner with overlaps, and a local count was then assigned to each patch by the network. Inferred redundant local counts were finally normalized and fused to the global count. Stahl *et al.* [41] regressed the counts for object proposals generated by Selective Search [45] and combined local counts using an inclusion-exclusion principle. Inspired by subitizing, the ability for a human to quickly counting a few objects at a glance, Chattopadhyay *et al.* [4] transferred their focus to the problem of counting objects in everyday scenes. The main challenge thus shifted to large intra-class variances rather than the occlusions and perspective distortions in crowded scenes.

While some methods above [4], [41] leverage the idea of spatial divisions, they still regress the open-set counts. Despite the fact that local region patterns are easier to be modelled than the whole image, the observed local patches are still limited. Since only finite local patterns (a closed set) can be observed, new scenes in reality have a high probability including objects out of the range (an open set). Moreover, dense regions with large count values are rare (Fig. 2) and the networks may suffer from sample imbalance. In this paper, we show that *a counting network is able to learn from a closed set with a certain range of counts, e.g., 0 ~ 20, and then generalizes to an open set (including counts > 20) via S-DC*.

2.3 Beyond Simple Regression

Regression is a natural approach to estimate continuous variables, such as age, depth, and counts. Some works suggest that regression is encouraged to be reformulated as an ordinal regression problem or a classification problem, which often enhances performance and benefits optimization [5], [11], [20], [27], [23] for many vision tasks. Ordinal regression is usually implemented by modifying well-studied classification algorithms and has been applied to the problem of age estimation [27] and monocular depth prediction [11]. Li *et al.* [20] further showed that directly reformulating regression to classification was also a good choice. In counting, the idea of blockwise classification is also investigated [23]. All these attempts motivate us to devise a classification-based closed-set counter. In this work, in addition to the standard regression-based modeling as in [25], SS-DCNet also follows [20] and [23] to discretize local counts and classify count intervals. Indeed, we observe in experiments that classification with S-DC generally works better than regression.

2.4 Open-Set Problems in Computer Vision

Many vision tasks are open-set by nature, such as depth prediction [11], [20], age estimation [5], [27], object recognition [35], visual domain adaptation [30], etc. While the sense of the open set may be different, they generally suffer from poor generalization as object counting. However, we find that, the learning target of counting alone, i.e., the count value, can be easily transformed into a closed set (via spatial division).

3 SUPERVISED SPATIAL DIVIDE-AND-CONQUER NETWORK

In this section, we describe how to construct a closed-set counter. We also explain our proposed SS-DCNet in detail.

3.1 Closed-Set Counter

In local count modeling, there are two approaches to define a counter in the closed set $[0, C_{max}]$, i.e., counting by regression [9], [25] and counting by classification [47], [23]. In practice, C_{max} should not be greater than the maximum local count observed in the training set. It is clear that treating C_{max} as the maximum prediction will cause a systematic error, but the error can be mitigated via S-DC, as discussed in Section 6.

Regression-Based Counter (R-Counter): R-Counter directly regresses count values within the closed set. If predicted count values are greater than C_{max} , the predictions will simply be truncated to C_{max} .

Classification-Based Counter (C-Counter): Instead of regressing open-set count values, C-Counter discretizes local counts and classifies count intervals as in [23]. Specifically, we define an interval partition of $[0, +\infty)$ as $\{0\}, (0, C_1], (C_2, C_3], \dots, (C_{M-1}, C_{max}]$ and $(C_{max}, +\infty)$. These $M + 1$ sub-intervals are labeled to the 0-th to the M -th classes, respectively. For example, if a count value falls into $(C_2, C_3]$, it is labeled as the 2-nd class. The median of each sub-interval can be adopted when recovering the count from the interval. Notice

Table 1

The Configurations of *Counter*, *Division decider* and *Upsampler*. *AvgPool* denotes Average Pooling. Convolutional layers are defined in the format: *kernel size Conv, output channel, s stride*. Each convolutional layer is followed by ReLU except the last layer. In particular, a *Sigmoid* function is attached at the end of *division decider* to generate soft division masks. A *Spatial Softmax* function is applied at the End of *Upsampler*, which constrains the sum of upsampling weights in each 2×2 adjacent regions to be 1 and ensures consistent local count values in the same image area after upsampling. The final output channel is 1 for R-Counter and *class num* for C-Counter

| Counter | Division decider/Upsampler |
|---------------------------------------|-----------------------------|
| 2×2 AvgPool, s 2 | 2×2 AvgPool, s 2 |
| 1×1 Conv, 512, s 1 | 1×1 Conv, 512, s 1 |
| 1×1 Conv, 1/(class num), s 1 | 1×1 Conv, 1, s 1 |
| — | Sigmoid/Spatial Softmax |

that, for the last sub-interval $(C_{max}, +\infty]$, C_{max} will be used as the count value if a region is classified into this interval.

In what follows, we term the network SS-DCNet (reg) when R-Counter is adopted, and SS-DCNet (cls) when C-Counter is used.

3.2 Single-Stage Spatial Divide-and-Conquer

As shown in Fig. 4, SS-DCNet includes a VGG16 [38] feature encoder, an UNet [34]-like decoder, a closed-set counter, a division decider and an upsampler. The counter, the structures of division decider and the upsampler are shown in Table 1. Note that, the first average pooling layer in the counter has a stride of 2, so the final prediction has an output stride of 64.

The feature encoder removes fully-connected layers from the pre-trained VGG16. Suppose that the input patch is of size 64×64 . Given the feature map F_0 (extracted from the Conv5 layer) with $\frac{1}{32}$ resolution of the input image, the counter predicts the local count value C_0 conditioned on F_0 . Note that C_0 is the local count without S-DC, which is also the final output of previous approaches [4], [9], [25].

We execute the first-stage S-DC on the fused feature map F_1 . F_1 is divided and sent to the shared counter to produce the division count $C_1 \in \mathbb{R}^{2 \times 2}$. Concretely, F_0 is upsampled by $\times 2$ in an UNet-like manner to F_1 . Given F_1 , the counter fetches the local features that correspond to spatially divided sub-regions, and predicts the first-level division counts C_1 . Each of the 2×2 elements in C_1 denotes a sub-count of the corresponding 32×32 sub-region.

With local counts C_0 and C_1 , the next question is to decide where to divide. We learn such decisions with another network module, division decider, as shown in the right part of Fig. 4. At the first stage of S-DC, the division decider generates a soft division mask W_1 of the same size as C_1 conditioned on F_1 such that for any $w \in W_1, w \in [0, 1]$. $w = 0$ means no division is required at this position, and the value in C_0 is used. $w = 1$ implies that here the initial prediction should be replaced with the division count in C_1 . Since both W_1 and C_1 are 2 times larger than C_0 , C_0 is required to be upsampled by $\times 2$ to \hat{C}_0 .

Note that, since C_0 denotes the local count of a 64×64 region, the sum of \hat{C}_0 should equal to C_0 . The upsampling of C_0 is therefore a re-distribution operator that assigns C_0 to each sub-region. We compute the re-distribution map U_1

Algorithm 1: Multi-Stage S-DC

Input: Image I and division time N
Output: Image count C

- 1 Extract F_0 from I ;
- 2 Generate C_0 given F_0 with the closed-set counter;
- 3 Initialize $DIV_0 = C_0$;
- 4 **for** $i \leftarrow 1$ to N **do**
- 5 Decode F_{i-1} to F_i ;
- 6 Process F_i with the closed-set counter, upsampler and the division decider to obtain C_i, U_i and the division mask W_i ;
- 7 Upsample DIV_{i-1} to \hat{DIV}_{i-1} with U_i as Eq. (3);
- 8 Update DIV_i as Eq. (4);
- 9 Integrate over DIV_N to obtain the image count C ;
- 10 **return** C

from the upsampler conditioned on F_1 , and the sum of U_1 equals to 1. We then upsample C_0 to \hat{C}_0 by

$$\hat{C}_0 = (C_0 \otimes \mathbf{1}_{2 \times 2}) \circ U_1, \quad (1)$$

where “ \otimes ” denotes Kronecker product and $\mathbf{1}_{2 \times 2}$ denotes a 2×2 matrix filled with 1. Finally, the first-stage division result DIV_1 takes the form

$$DIV_1 = (\mathbf{1} - W_1) \circ \hat{C}_0 + W_1 \circ C_1, \quad (2)$$

where $\mathbf{1}$ denotes a matrix filled with 1 and is with the same size of W_1 , and “ \circ ” denotes the Hadamard product.

3.3 Multi-Stage Spatial Divide-and-Conquer

SS-DCNet can execute multi-stage S-DC by further decoding, dividing the feature map until reaching the output of the first convolutional block. In this sense, the maximum division time is 4 in VGG16 for example. Actually we show later in experiments that a two-stage division is sufficient to achieve satisfactory performance. In multi-stage S-DC, DIV_{i-1} ($i \geq 2$) is first upsampled as:

$$\hat{DIV}_{i-1} = (DIV_{i-1} \otimes \mathbf{1}_{2 \times 2}) \circ U_i, \quad (3)$$

and then merged according to

$$DIV_i = (\mathbf{1} - W_i) \circ \hat{DIV}_{i-1} + W_i \circ C_i, \quad (4)$$

in a recursive manner. Multi-stage SS-DCNet is summarized in Algorithm 1.

3.4 Loss Functions

Here we elaborate the loss functions used in an N -stage SS-DCNet.

Counter Loss: As mentioned in Section 3.1, both R-Counter and C-Counter can be used. We use the ℓ_1 loss, denoted by L_R^i , $i = 0, 1, 2, \dots, N$, for each level of output C_i when R-Counter is used, and cross-entropy loss, denoted by L_C^i , $i = 0, 1, 2, \dots, N$, when C-Counter is chosen. Note that, both ground-truth local counts and predicted counts are truncated to C_{max} when R-Counter is adopted. The overall counter loss is $L_R = \sum_{i=0}^N L_R^i$ for the R-Counter and $L_C = \sum_{i=0}^N L_C^i$ for the C-Counter.

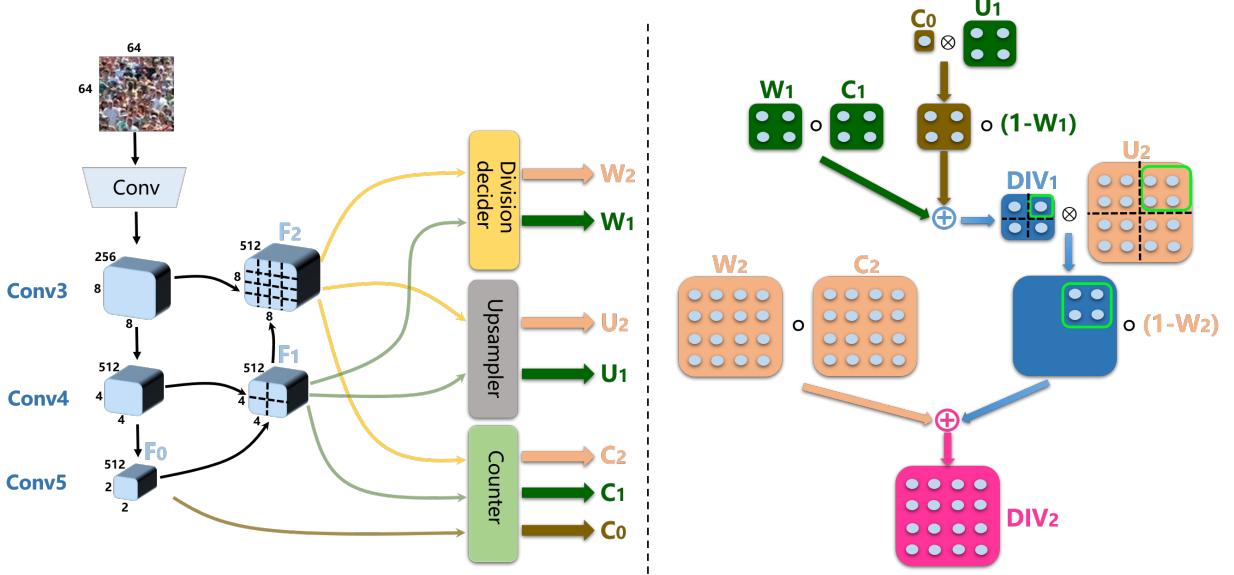


Figure 4. The architecture of SS-DCNet (left) and a two-stage S-DC process (right). SS-DCNet adopts all convolutional layers in VGG16 [38] (the first two convolutional blocks are simplified as *Conv*). An UNet [34]-like decoder is employed to upsample and divide the feature map as in Fig. 3. A shared upsampler, a closed-set counter and a division decider receive divided feature maps, and respectively, generate upsampling map U_i s, division counts C_i s and division masks W_i s, for $i = 1, 2, \dots$. After obtaining these results, C_{i-1} is upsampled with U_i , then it is merged with C_i by W_i to the i -th division count DIV_i shown in the right sub-figure. In particular, we upsample each count of low resolution into the corresponding 2×2 area of high resolution before merging with U_i . “ \circ ” denotes the Hadamard product, and “ \otimes ” denotes the Kronecker product. Note that, the 64×64 local patch is only used as an example for readers to understand the pipeline of SS-DCNet. Since SS-DCNet is a fully convolutional network, it can process an image of arbitrary sizes, say $M \times N$, and return DIV_2 of size $\frac{M}{64} \times \frac{N}{64}$. The configurations of the closed-set counter and the division decider are presented in Table 1.

Merging Loss (Implicit Division Supervision): We also adopt a ℓ_1 loss L_m for the final division output DIV_N . L_m provides an implicit supervision signal for learning W_i s.

Division Loss (Explicit Division Supervision): We can also explicitly supervise W_i by comparing the ground-truth C_i^{gt} and C_{max} . As shown in Fig. 5, if the ground-truth count value of a 64×64 local region D , i.e., C_0^{gt} , is larger than C_{max} , the inferred count C_0 will be no larger than C_{max} . Let us assume $C_0 = C_{max}$. One knows that this local region is under-estimated ($C_{max} < C_0^{gt}$), but it does not imply that all sub-regions of D are underestimated. As shown in Fig. 5, there are 4 possibilities where underestimations occur. In this case, we can only know that at least one sub-region of D is underestimated and is required to be replaced with C_1 , which means at least one of values of W_1 should approach 1. Hence, we constrain this value to be the one with the largest probability approaching 1 (the maximum) when $C_0^{gt} > C_{max}$.

$$L_{div}^1 = -\mathbb{1}\{C_0^{gt} > C_{max}\} \times \log(\max(W_1)), \quad (5)$$

where $\mathbb{1}\{P\}$ denotes the indicator function which outputs 1 when the condition P is true, and 0 otherwise. C_0^{gt} is the ground truth count value of C_0 , and ‘max’ is the operator that returns the maximum value. Following Eq. (5), the loss L_{div}^i , $i = 1, 2, \dots, N$, for W_i can be deduced as

$$L_{div}^i = -\sum_{j=1}^{H_i} \sum_{k=1}^{W_i} \mathbb{1}\{C_{i-1}^{gt}[j, k] > C_{max}\} \times \log(\max(W_i[2j-1 : 2j, 2k-1 : 2k])), \quad (6)$$

where $C_{i-1}^{gt}[j, k]$ denotes the element of the j -th row and the k -th column of C_{i-1}^{gt} , and $W_i[2j-1 : 2j, 2k-1 : 2k]$ the

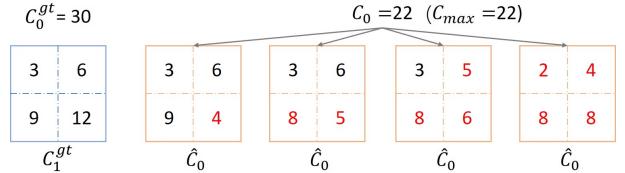


Figure 5. Motivation of the division loss. When $C_0^{gt} > C_{max}$, the prediction C_0 can only be C_{max} at most, and it is sure that at least one quarter of the region is underestimate.

avg: average upsampling

\times : Kronecker product

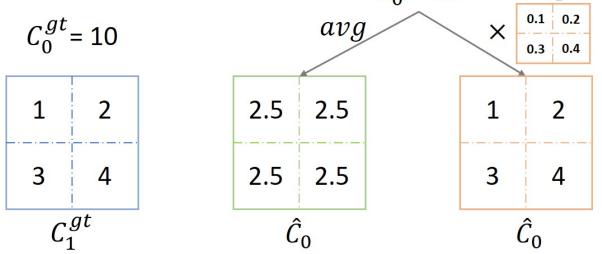


Figure 6. Motivation of the upsampling loss. The middle depicts averaging upsampling in S-DCNet [47], and the right shows guided upsampling in SS-DCNet informed by the upsampling map U_1 . Locally accurate U_1 leads to locally accurate \hat{C}_0 .

elements lying in the $(2j-1)$ -th to $2j$ -th row, $(2k-1)$ -th to $2k$ -th column of W_i . The overall division loss $L_{div} = \sum_{i=1}^N L_{div}^i$.

Upsampling Loss: By comparing count upsampling between S-DCNet [47] and SS-DCNet in Fig. 6, we find that, even if C_0 is accurately predicted, average upsampling in S-DCNet can

produce inaccurate \hat{C}_0 . To alleviate such errors, we introduce an upsampling map U_1 to guide the upsampling of C_0 . An upsampling loss is thus required to supervise the learning of U_1 s.

The upsampler predicts U_i used to re-distribute DIV_{i-1} to its $\times 2$ resolution output \hat{DIV}_{i-1} . We compute the ground truth of U_i according to the distribution of local counts:

$$U_i^{gt} = C_i^{gt} / (C_{i-1}^{gt} \otimes \mathbf{1}_{2 \times 2}), \quad (7)$$

where U_i^{gt} , C_i^{gt} and C_{i-1}^{gt} denote the ground truth of U_i , C_i and C_{i-1} , respectively. $'/'$ denotes element-wise division. We use ℓ_1 loss L_{up}^i , $i = 1, 2, \dots, N$, for each U_i . Hence, the overall upsampling loss $L_{up} = \sum_{i=1}^N L_{up}^i$.

Division Consistency Loss: When R-Counter is used, we can further constrain the consistency between different C_i s when C_i s are in the range of the closed set $[0, C_{max}]$, which shares a similar spirit compared to [36]. For C_0 and C_1 , the division consistency loss is defined by

$$L_{eq}^1 = \mathbb{1}\{C_0^{gt} \leq C_{max}\} \times |C_0 - \text{sum}(C_1)|, \quad (8)$$

where ‘sum’ is the operator that returns the sum of all elements. Following Eq. 8, the consistency loss between C_{i-1} and C_i , $i = 1, 2, \dots, N$, is

$$L_{eq}^i = \sum_{j=1}^{H_{i-1}} \sum_{k=1}^{W_{i-1}} \mathbb{1}\{C_{i-1}^{gt}[j, k] \leq C_{max}\} \times |C_{i-1}[j, k] - \text{sum}(C_i[2j-1 : 2j, 2k-1 : 2k])| \quad . \quad (9)$$

The overall division consistency loss $L_{eq} = \sum_{i=1}^N L_{eq}^i$. Note that, when C-Counter is adopted, the gradient of the consistency loss cannot be back propagated because count values are discretized into count intervals and represented by class labels. We simply drop the consistency loss in this case.

As a summary, for SS-DCNet (reg) the final loss L_{reg} is

$$L_{reg} = L_R + L_m + L_{up} + L_{div} + L_{eq}, \quad (10)$$

and for SS-DCNet (cls) the final loss L_{cls} is

$$L_{cls} = L_C + L_m + L_{up} + L_{div}. \quad (11)$$

4 OPEN SET OR CLOSED SET? A THEORETICAL ANALYSIS

How does SS-DCNet benefit from transforming count values from open set to closed set? Here we first give the mathematical definitions of the open set, the closed set and the spatial division. With these definitions, we attempt to answer how many division times are required for transferring counts from the open set to the closed set in Proposition 1. Then in Proposition 2, we show that, with sufficient spatial divisions, transforming count values from the open set can lead to lower absolute errors on the closed set, which sheds light on why we model counts in a closed set.

Definition 1 (Spatial Division of an Image). *Given an image $I \in \mathbb{R}^{H \times W \times K}$, where H, W and K denote the height, width and channel dimensions, respectively, the spatial division of I leads to a group of sub-images $\{I_i \in \mathbb{R}^{H_i \times W_i \times K}\}_{i=1,2,\dots,M}$ that satisfy: i) $I_i \subset I$;*

- ii) $I_i \cap I_j = \emptyset$, for $i \neq j$;
- iii) $I_1 \cup I_2 \cup \dots \cup I_M = I$.

Definition 2 (Open Set and Closed Set). *Given a positive number C_{max} , for $\forall x \geq 0$ and $x \in \mathbb{R}$, we can define a closed set \mathcal{S}_C by $\{x | 0 \leq x \leq C_{max}\}$ and an open set \mathcal{S}_O by $\{x | x > C, C \geq C_{max}\}$.*

In object counting, C_{max} is the maximum count value observed in the training set. Note that here we define the open set \mathcal{S}_O to be $(C_{max}, +\infty]$, rather than $[0, +\infty]$ aforementioned. This is because $[0, +\infty)$ and \mathcal{S}_C only differ in the range of $(C_{max}, +\infty]$, where S-DC is applied to this range to transform count values into \mathcal{S}_C . The shared interval $[0, C_{max}]$ remains unchanged and does not require S-DC. Hence, we define \mathcal{S}_O to be disjoint from \mathcal{S}_C to simplify the analysis.

Lemma 1. *Given an image I , let \boxplus be the spatial dichotomy division operator such that $\boxplus(I) = \{I_i\}_{i=1,2,3,4}$. Let \boxplus^N further denote \boxplus is applied for N times. We have $\boxplus^N(I) = \{I_j\}_{j=1,2,\dots,4^N}$.*

Proof. Suppose that M is the number of divided sub-images after N divisions.

i) For $N = 1$, according to the definition of \boxplus ,

$$\boxplus(I) = \{I_i\}_{i=1,2,3,4}, \quad (12)$$

which means $M = 4$;

ii) For $N = t$, assume $M = 4^t$, we have

$$\boxplus^t(I) = \{I_k\}_{k=4^t}, \quad (13)$$

then when $N = t + 1$, for each $I_k \in \boxplus^t(I)$,

$$\begin{aligned} \boxplus^{(t+1)}(I) &= \boxplus(\boxplus^t(I)) \\ &= \{\boxplus(I_k), k = 1, 2, \dots, 4^t\}, \\ &= \{I_{kp}\}_{k=1,2,\dots,4^t, p=1,2,3,4} \end{aligned} \quad (14)$$

so $M = 4 \times 4^t = 4^{t+1}$ holds for $N = t + 1$.

Since both i) and ii) hold, by mathematical induction, we can deduce $M = 4^N$ after N divisions. \square

According to Lemma 1, we know that, an image I will be divided into 4^N sub-images at most after N divisions. A subsequent question of interest is that, *how many spatial divisions are required to transfer count values from \mathcal{S}_O to \mathcal{S}_C ?* This leads to our following proposition.

Proposition 1 (Minimum and Maximum Division Times). *Assume an image $I \in \mathbb{R}^{H \times W \times K}$ with a count value $C^* > C_{max}$, $C^* \in \mathcal{S}_O$, is divided by the \boxplus operator, and $r \times r$ is the minimum sub-region size with C_{max} objects, then the required division times N for transferring C^* into \mathcal{S}_C satisfy*

$$\left\lceil \log_4 \frac{C^*}{C_{max}} \right\rceil \leq N \leq \left\lceil \max\{\log_2 \frac{H}{r}, \log_2 \frac{W}{r}\} \right\rceil + 1.$$

Proof. Suppose after N division times, I is divided into M sub-images $\{I_i \in \mathbb{R}^{H_i \times W_i \times K}\}_{i=1,2,\dots,M}$ with local count values c_i s that satisfy $0 \leq c_i \leq C_{max}$ and $\sum_{i=1}^M c_i = C^*$.

i) Minimum Division Times. Since

$$c_i \leq \max_{i=1,2,\dots,M} c_i \leq C_{max}, \quad (15)$$

we have

$$C^* = \sum_{i=1}^M c_i \leq M \times \max_{i=1,2,\dots,M} c_i \leq M \times C_{\max}. \quad (16)$$

Hence, $M \geq \frac{C^*}{C_{\max}}$. With Lemma 1, we know $M = 4^N$, so $4^N \geq \frac{C^*}{C_{\max}}$, i.e., $N \geq \log_4 \frac{C^*}{C_{\max}}$.

Note that, since N is an integer, the minimum division times are $\lceil \log_4 \frac{C^*}{C_{\max}} \rceil$.

ii) Maximum Division Times. First, we state that, if the size of all sub-images I_i is less than $r \times r$ ($H_i < r$ and $W_i < r$ for $i = 1, 2, \dots, M$), then the counts c_i s of these sub-images will be less than C_{\max} , i.e., $\max_{i=1,2,\dots,M} c_i < C_{\max}$.

We prove this statement with proof by contradiction. Suppose that there exists a sub-image I_i of size $H_i \times W_i$ containing c_i objects with $c_i \geq C_{\max}$. Since I_i has no less than C_{\max} objects, the minimum region size of C_{\max} objects cannot exceed the size of I_i , i.e., $r \leq \max\{H_i, W_i\}$. This contradicts with the assumption that $H_i < r$ and $W_i < r$. Hence, the assumption does not hold.

Let N_r denote the division times required to ensure that the sizes of all sub-images I_i s are no larger than $r \times r$, i.e., N_r satisfies

$$\max\left\{\frac{H}{2^{N_r}}, \frac{W}{2^{N_r}}\right\} < r. \quad (17)$$

Since $N_r > \max\{\log_2 \frac{H}{r}, \log_2 \frac{W}{r}\}$ and N_r is an integer, we have

$$N_r = \left\lfloor \max\{\log_2 \frac{H}{r}, \log_2 \frac{W}{r}\} \right\rfloor + 1. \quad (18)$$

Hence, the required division times $N \leq N_r$.

Proof completes. \square

Proposition 1 suggests the lower and higher bounds when SS-DCNet can transform count values from \mathcal{S}_O to \mathcal{S}_C . This is a prerequisite that SS-DCNet can work. Proposition 1 also allows one to have a prior estimate of the degree of granularity required in the spatial divisions. Is such a transformation effective? We further provide a theoretical analysis based on the metric of the absolute error. It is worth noting that the absolute error is widely considered to be an evaluation metric for counting, and we analyse the absolute counting error of the closed set and the open set with the help of the relative error. This is because that only the relative error normalized by the ground truth count can link the counting error across a wide range and provide a fair comparison between two distinct sets.

Definition 3. Let C ($C > 0$) be the ground-truth value, and \hat{C} the inferred value. We define the relative error by $\epsilon_r = \frac{C - \hat{C}}{C}$ and the absolute error by $\epsilon_a = |C - \hat{C}| = C \times |\epsilon_r|$.

By Definition 3, it is clear that the expectation of $|\epsilon_r|$ varies as the ground truth C changes. We thus have

Definition 4. The function of the expectation w.r.t. $|\epsilon_r|$ is defined by $f(x) = \mathbb{E}_{C=x} |\epsilon_r| = \mathbb{E}_{C=x} \left| \frac{C - \hat{C}}{C} \right|$, and $f(x)$ is assumed to be continuous.

Before presenting our main results, we further need the conclusion of the following theorem.

Theorem 1 (Extreme Value Theorem [32]). If $f(x)$ is a continuous function defined in the closed interval $[a, b]$, then $\exists c \in [a, b]$ that satisfies $f(c) = \max_{a \leq x \leq b} f(x)$.

According to Definition 3, Definition 4 and Theorem 1, we arrive at our final proposition.

Proposition 2. Let ϵ_a^o denote the absolute counting error on \mathcal{S}_O , ϵ_a^c the absolute counting error on \mathcal{S}_C (after sufficient spatial divisions as in Proposition 1), $f(x) = \mathbb{E}_{C=x} |\epsilon_r|$, and C_{\max} a predefined positive number. Given a count value C^* , if $C^* > C_{\max}$ and $f(C^*) > \max_{0 \leq x \leq C_{\max}} f(x)$, then

$$\mathbb{E}_{C=C^*} [\epsilon_a^c] \leq \max_{0 \leq x \leq C_{\max}} f(x) \times C^* < \mathbb{E}_{C=C^*} [\epsilon_a^o].$$

Proof. For an image I with a count value C^* , a spatial division of I , i.e., $\{I_i\}_{i=1,2,\dots,M}$, could be found in Definition 1. Their corresponding local counts $\{c_i\}_{i=1,2,\dots,M}$ satisfy i) $0 \leq c_i \leq C_{\max}$, $i = 1, 2, \dots, M$, and ii) $\sum_{i=1}^M c_i = C^*$.

Let the ϵ_r^o denote the relative counting error on \mathcal{S}_O , and ϵ_r^i the relative counting error of each I_i on \mathcal{S}_C . By Definition 3, we have

$$\epsilon_a^o = |C^* \times \epsilon_r^o| = C^* \times |\epsilon_r^o|, \quad (19)$$

and

$$\epsilon_a^c = \left| \sum_{i=1}^M c_i \times \epsilon_r^i \right|. \quad (20)$$

By Definition 4,

$$\begin{aligned} \mathbb{E}_{C=C^*} [\epsilon_a^o] &= \mathbb{E}_{C=C^*} [C^* \times |\epsilon_r^o|] \\ &= C^* \times \mathbb{E}_{C=C^*} |\epsilon_r^o| \\ &= C^* \times f(C^*) \end{aligned} \quad (21)$$

With generalized triangle inequality [17], we have

$$\epsilon_a^c = \left| \sum_{i=1}^M c_i \times \epsilon_r^i \right| \leq \sum_{i=1}^M |c_i \times \epsilon_r^i| = \sum_{i=1}^M c_i \times |\epsilon_r^i|. \quad (22)$$

By taking the expectation of both sides of Eq. (22), it amounts to

$$\begin{aligned} \mathbb{E}_{C=C^*} [\epsilon_a^c] &\leq \sum_{i=1}^M c_i \times \mathbb{E}_{C=C^*} |\epsilon_r^i| \\ &= \sum_{i=1}^M c_i \times f(c_i) \end{aligned} \quad . \quad (23)$$

According to Theorem 1, $\exists \zeta \in [0, C_{\max}]$ such that $f(\zeta) = \max_{0 \leq x \leq C_{\max}} f(x)$. Hence, $f(c_i) \leq f(\zeta)$, for $i = 1, 2, \dots, M$, so

$$\begin{aligned} \mathbb{E}_{C=C^*} [\epsilon_a^c] &\leq \sum_{i=1}^M c_i \times f(c_i) \\ &\leq \sum_{i=1}^M c_i \times f(\zeta) \\ &= f(\zeta) \times \sum_{i=1}^M c_i \\ &= f(\zeta) \times C^* \end{aligned} \quad . \quad (24)$$

If $f(C^*) > f(\zeta) = \max_{0 \leq x \leq C_{max}} f(x)$, with Eq. (21), we have

$$\mathbb{E}_{C=C^*}[\epsilon_a^c] \leq f(\zeta) \times C^* < f(C^*) \times C^* = \mathbb{E}_{C=C^*}[\epsilon_a^o]. \quad (25)$$

Proof completes. \square

Proposition 2 states that, by transforming the count value from \mathcal{S}_O to \mathcal{S}_C , SS-DCNet can achieve lower counting errors on the condition that the expectation of the relative errors on \mathcal{S}_C is smaller than that on \mathcal{S}_O . We will verify this condition via experiments in Section 5. According to Proposition 2, it is encouraged to model counting in a closed set in theory.

It is worth noting that, although our theory is developed specifically for object counting, the theoretical results are generic and not limited to this task. As long as the learning target is spatially divisible as the count value (so far we only find counting satisfies the property of spatial divisibility), without loss of generality, the same conclusion can be deduced, as stated in Corollary 1.

Definition 5 (Spatial Divisibility). Let $P \in [0, +\infty)$ be the learning target of a vision task defined on an Image I . $\{I_i\}_{i=1,2,\dots,M}$ is the spatial division of I , and p_i is the corresponding learning target of each I_i . P is spatially divisible if $\sum_{i=1}^M p_i = P$.

Corollary 1. Let P be the learning target and is spatially divisible. Let ϵ_a^o denote the absolute error of P on \mathcal{S}_O , ϵ_a^c the absolute error of P on \mathcal{S}_C (after spatial divisions), $f(x) = \mathbb{E}_{P=x}|\epsilon_r|$, and P_{max} a predefined positive number. Given a positive number P^* , if $P^* > P_{max}$ and $f(P^*) > \max_{0 \leq x \leq P_{max}} f(x)$, then

$$\mathbb{E}_{P=P^*}[\epsilon_a^c] \leq \max_{0 \leq x \leq P_{max}} f(x) \times P^* < \mathbb{E}_{P=P^*}[\epsilon_a^o].$$

5 OPEN SET OR CLOSED SET? A JUSTIFICATION ON A SYNTHETIC DATASET

As aforementioned, counting is an open-set problem, while the model is learned in a closed set. *Can a closed-set counting model really generalize to open-set scenarios?* Here we show through a controlled toy experiment that, the answer is negative. In addition, in this experiment we illustrate that SS-DCNet indeed works better than that without S-DC, which supports our Proposition 2. Inspired by [18], we synthesize a cell counting dataset to explore the counting performance outside a closed training set.

5.1 Synthetic Cell Counting Dataset

We first generate 500 256×256 images with 64×64 sub-regions containing only $0 \sim 10$ cells to construct the training set (a closed set). To generate an open testing set, we further synthesize 500 images with sub-region counts uniformly distributed in the range of $[0, 20]$.

5.2 Baselines and Protocols

We implement three approaches for comparisons, which are: *i*) a density regression baseline CSRNet [21]; *ii*) a regression baseline with pretrained VGG16 as the backbone and the R-Counter used in SS-DCNet as the backend, without S-DC. ℓ_1 loss is used. This baseline directly regresses the open-set counts; *iii*) a classification baseline with the same VGG16 and the C-Counter, without S-DC; *iv*) our proposed SS-DCNet, which learns from a closed set but adapts to the open set via S-DC. According to Proposition 1, at least 1-time division is required for SS-DCNet to transform count values from the open set to the closed set. We adopt both SS-DCNet (reg) and SS-DCNet (cls) with 1-time division for comparison.

As for the discretization of count intervals, we choose 0.5 as the step because cells may be overlapping in local patches. Hence, we have a partition of $\{0\}, (0, 0.5], (0.5, 1], \dots, (9.5, 10]$ and $(10, +\infty)$. All approaches are trained with standard stochastic gradient descent (SGD). The learning rate is initially set to 0.0001 and is decreased by $\times 10$ when the training error stagnates.

5.3 Observations

According to Fig. 7(b), it can be observed that both regression and classification baselines work well in the range of the closed set ($0 \sim 10$), but the counting error increases quickly when counts are larger than 10. This suggests that a conventional counting model learned in a closed set cannot generalize to the open set. However, SS-DCNet can achieve accurate predictions even on the open set, which confirms the advantage of S-DC.

5.4 Analyses

The relative mean absolute error (rMAE) is an empirical estimate of $f(x)$ according to Definition 4, and the MAE is also an empirical estimate of $\mathbb{E}_{C=C_0}\{\epsilon_a\}$. Fig. 7(c) and (b) report how these two metrics vary, respectively. We have the following discussions:

- As shown in Fig. 7(c), $f(C_0) > \max_{0 \leq x \leq C_{max}} f(x)$ satisfies for the classification baseline when the patch count $C_0 \geq 12$. According to Proposition 2, under the condition above, SS-DCNet (cls) will show lower MAE than the classification baseline without S-DC. When $C_0 \geq 17$, the same conclusion can be drawn between SS-DCNet (reg) and its open-set regression baseline.
- When $C_0 \in [10, 12]$, $f(C_0) > \max_{0 \leq x \leq C_{max}} f(x)$ is no longer true for the classification baseline. As shown in Fig. 7(b), SS-DCNet (cls) only reports comparable results against the classification baseline. When $C_0 \in [10, 17]$, the same observation can be made between SS-DCNet (reg) and its open-set regression counterpart.

In general, our experiment on the synthetic data verifies Proposition 2 to some extent. According to these results, it is also encouraged to model counting in a closed set in practice.

6 EXPERIMENTS ON REAL DATASETS

Extensive experiments are further conducted to demonstrate the effectiveness of SS-DCNet on real datasets. We first describe some essential implementation details. Then ablation

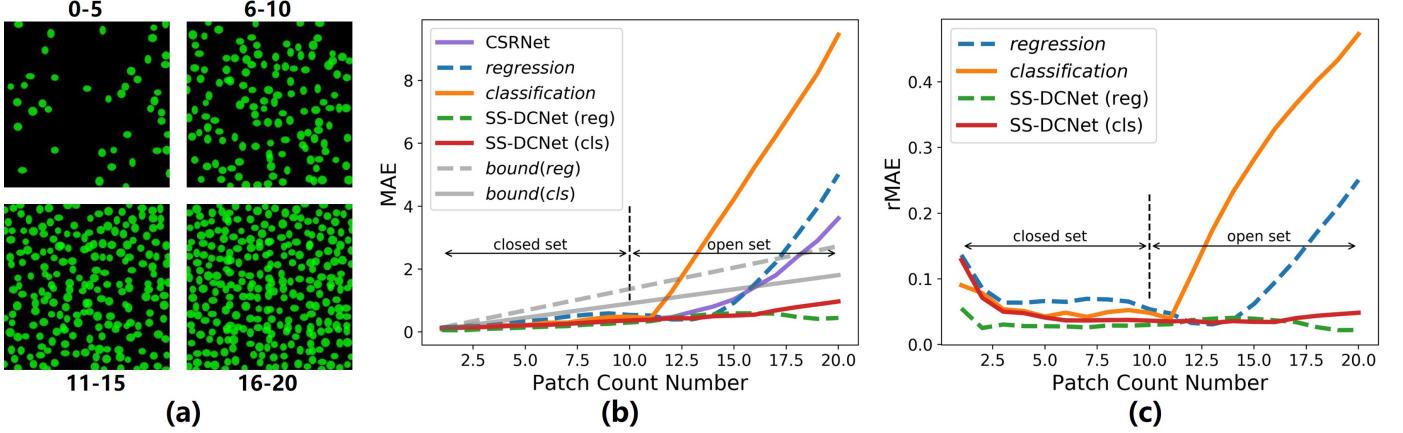


Figure 7. A toy-level justification. (a) Some 256×256 images in the simulated cell counting dataset. The numbers denote the range of local counts of 64×64 sub-regions. (b) The mean absolute error (MAE) of different methods with increased 64×64 sub-region counts. (c) The relative mean absolute error (rMAE) of different methods with increased 64×64 sub-region counts. SS-DCNet (reg/cls) adopts one-stage S-DC.

studies are conducted on the ShanghaiTech Part_A [51] dataset to highlight the benefit of S-DC. We then compare SS-DCNet against current state-of-the-art methods on five public datasets. Finally, we also report cross-domain performance to verify the generalization ability of SS-DCNet.

Mean Absolute Error (MAE) and Root Mean Squared Error (MSE) are chosen to quantify the counting performance. They are defined by

$$MAE = \frac{1}{Z} \sum_{i=1}^Z |C_i^{pre} - C_i^{gt}|, \quad (26)$$

$$MSE = \sqrt{\frac{1}{Z} \sum_{i=1}^Z (C_i^{pre} - C_i^{gt})^2}, \quad (27)$$

where Z denotes the number of images, C_i^{pre} denotes the predicted count of the i -th image, and C_i^{gt} denotes the corresponding ground-truth count. MAE measures the accuracy of counting, and MSE measures the stability. Lower MAE and MSE imply better counting performance.

In addition, the absolute error is not always meaningful, because a mistake of 1 for a ground truth count of 2 might seem egregious but the same mistake for the ground truth count of 23 might seem reasonable. This is rooted in the fact that human perception of count is essentially logarithmic and not linear [10]. Aside from MAE and MSE , we further report the relative Mean Absolute Error (rMAE) for most datasets used, defined by

$$rMAE = \frac{1}{Z} \sum_{i=1}^Z \frac{|C_i^{pre} - C_i^{gt}|}{C_i^{gt}}, \quad (28)$$

6.1 Implementation Details

6.1.1 Interval Partition for C-Counter

We generate ground-truth counts of local patches by integrating over the density maps. The counts are usually not integers, because objects can partly present in cropped local patches. We evaluate two different partition strategies. In the first partition, we choose 0.5 as the step and generate partitions as $\{0\}, (0, 0.5], (0.5, 1], \dots, (C_{max} - 0.5, C_{max}]$ and $(C_{max}, +\infty)$, where C_{max} denotes the maximum count

of the closed set. This partition is named as One-Linear Partition.

In the second partition, we further finely divide the sub-interval $(0, 0.5]$, because this interval contains a sudden change from no object to part of an object, and a large proportion of objects lie in this sub-interval. A small step of 0.05 is further used to divide the sub-interval $(0, 0.5]$, i.e., $(0, 0.05], (0.05, 0.1], \dots, (0.45, 0.5]$. Other intervals remain the same as One-Linear Partition. We call this partition Two-Linear Partition.

6.1.2 Data Preprocessing

We follow the same data augmentation used in [21], except for the UCF-QNRF dataset [15] where we adopt two data augmentation strategies. In particular, 9 sub-images of $\frac{1}{4}$ resolution are cropped from the original image. The first 4 sub-images are from four corners, and the remaining 5 are randomly cropped. Random scaling and flipping are also executed.

6.1.3 Training Details

SS-DCNet is implemented with PyTorch [31]. We train SS-DCNet using SGD. The encoder in SS-DCNet is directly adopted from convolutional layers of VGG16 [38] pretrained on ImageNet, and the other layers employ random Gaussian initialization with a standard deviation of 0.01. The learning rate is initially set to 0.001 and is decreased by $\times 10$ when the training error stagnates. We keep training until convergence. For the ShanghaiTech, UCF_CC_50, TRANCOS and MTC datasets, the batch size is set to 1. For the UCF-QNRF dataset, the batch size is set to 16 following [15].

6.2 Ablation Study on the ShanghaiTech Part_A

6.2.1 Is SS-DCNet Robust to C_{max} ?

When reformulating the counting problem into classification, a critical issue is how to choose C_{max} , which defines the closed set. Hence, it is important that SS-DCNet is robust to the choice of C_{max} .

We conduct a statistical analysis on count values of local patches in the training set, and then set C_{max} with the quantiles ranging from 100% to 80% (decreased by

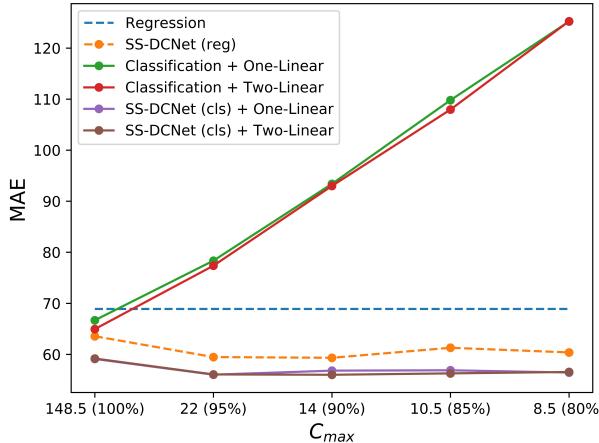


Figure 8. The influence of C_{max} to SS-DCNet on the ShanghaiTech Part_A dataset [51]. The numbers in the brackets denote quantiles of the training set, for example, 22 (95%) means the 95% quantile is 22. ‘VGG16 Encoder’ is the classification baseline without S-DC. ‘One-Linear’ and ‘Two-Linear’ are defined in Section 6.1.1. SS-DCNet (reg/cls) adopts two-stage S-DC.

Table 2
Results of SS-DCNet with different S-DC stages. The best performance is in boldface.

| Division time | SS-DCNet (cls) | | | SS-DCNet (reg) | | |
|---------------|----------------|-------------|---------------|----------------|-------------|---------------|
| | MAE | MSE | rMAE | MAE | MSE | rMAE |
| 0 | 76.0 | 142.5 | 16.48% | 76.7 | 144.6 | 16.41% |
| 1 | 57.8 | 92.0 | 13.81% | 61.0 | 98.1 | 14.68% |
| 2 | 56.1 | 88.9 | 13.78% | 59.5 | 95.0 | 14.35% |
| 3 | 57.0 | 92.7 | 13.98% | 60.1 | 97.2 | 14.37% |
| 4 | 59.1 | 100.0 | 14.23% | 62.8 | 99.1 | 15.39% |

5%). Two-stage SS-DCNet is evaluated. Another baseline of classification without S-DC is also used to explore whether counting can be simply modeled in a closed-set classification manner. To be specific, we reserve the VGG16 encoder and the C-Counter in this classification baseline.

Results are presented in Fig. 8. We see that the MAE of the classification baseline increases rapidly with decreased C_{max} . This result is not surprising, because the model is constrained to be visible to count values not greater than C_{max} . This suggests that counting cannot be simply transformed into closed-set classification. However, with the help of S-DC, SS-DCNet exhibits strong robustness to the changes of C_{max} . It seems that the systematic error brought by C_{max} can somewhat be alleviated with S-DC. As for how to choose a proper value for C_{max} , the maximum count of the training set seems not the best choice, while setting to some smaller values even delivers better performance. It may be due to that a model is only able to count objects accurately within a certain degree of denseness. We also notice that Two-Linear Partition is slightly better than One-Linear Partition, which indicates that the fine division to the (0, 0.5] sub-interval has a positive effect.

According to the results above, SS-DCNet is robust to C_{max} in a wide range of values, and C_{max} is generally encouraged to be set less than the maximum count value observed. In addition, there is no significant difference

between two kinds of partitions. For simplicity, we set C_{max} to be the 95% quantile and adopt Two-Linear Partition in the following experiments.

6.2.2 How Many Times to Divide?

SS-DCNet can apply S-DC by up to 4 times, but how many times are sufficient? Here we evaluate SS-DCNet with different division stages. The maximum count value of 64×64 image patches in the test set is 136.50 and $C_{max} = 22$. With Proposition 1, we know twice division is required at least. Quantitative results are listed in Table 2. It can be observed that when the division time N varies from 0 to 2, the counting error MAE and $rMAE$ significantly decreases for both SS-DCNet (reg) and SS-DCNet (cls). However, counting accuracy saturates when N continues increasing. In general, two-stage S-DC seems sufficient. We use this setup in the following experiments.

6.2.3 The Effect of S-DC

To highlight the effect of S-DC, we compare SS-DCNet against several regression and classification baselines. These baselines adopt the same architecture of VGG16 encoder and the counter in SS-DCNet. *classification* is the result of C_0 adopting C-Counter without S-DC, and C_{max} is set to be the 95% quantile ($C_{max} = 22$). For regression baselines, we employ R-counter to obtain the prediction C_0 without S-DC. We create two regression baselines. *open-set regression* + S-DC is straightforward. We do not limit the output range, and it can vary from 0 to $+\infty$. *S_c regression* indicates that the output range is constrained within $[0, C_{max}]$ (C_{max} is also set to 22 for a fair comparison). Any large outputs will be clipped to C_{max} .

Results are shown in Table 3. We can see that counting by classification without S-DC suffers from the limitation of C_{max} and performs even worse than S_o regression. S_c regression also suffers from the same problem. However, with S-DC, SS-DCNet (reg/cls) significantly reduces the counting error and outperforms both their regression/classification baseline by a large margin. It suggests that a counting model can learn from a closed set and generalize well to an open set via S-DC. We notice that SS-DCNet (cls) performs better than SS-DCNet (reg). It seems that reformulating counting in classification is more effective than in regression. One plausible reason is that the optimization is easier and less sensitive to sample imbalance in classification than in regression.

We further analyze the counting error of 64×64 local patches in detail. As shown in Fig. 9, we observe that the direct prediction C_0 without S-DC performs worse than the S_o regression baseline and CSRNet, which can be attributed to the limited C_{max} of the C-Counter. After embedding S-DC, the counting errors (MAE and $rMAE$) of DIV_1 and DIV_2 significantly reduce and outperform open-set regression and CSRNet. Such a benefit is even much clear in dense patches with local counts greater than 100. It justifies our argument that, instead of regressing a large count value directly, it is more accurate to count dense patches through S-DC, which verifies the conclusion in Proposition 2 in the real-world dataset.

Table 3
Effect of S-DC. The best performance is in boldface.

| Method | MAE | MSE | rMAE |
|---------------------------------|-------------|-------------|---------------|
| classification | 77.4 | 149.3 | 17.13% |
| <i>S_c</i> regression | 76.5 | 140.9 | 16.80% |
| <i>S_o</i> regression | 68.9 | 112.1 | 16.43% |
| SS-DCNet (reg) | 59.5 | 95.0 | 14.35% |
| SS-DCNet (cls) | 56.1 | 88.9 | 13.78% |

Table 4
Effect of different loss functions.

| Method | L_R | L_C | L_m | L_{up} | L_{div} | L_{eq} | MAE | MSE | rMAE |
|--------------------|-------|-------|-------|----------|-----------|----------|-------------|-------------|---------------|
| S-DCNet (reg) [47] | ✓ | | ✓ | | | | 64.7 | 105.7 | 15.84% |
| SS-DCNet (reg) | ✓ | | ✓ | ✓ | ✓ | | 61.5 | 99.2 | 15.34% |
| | ✓ | | ✓ | ✓ | ✓ | ✓ | 60.6 | 96.5 | 15.46% |
| | ✓ | | ✓ | ✓ | ✓ | ✓ | 59.5 | 95.0 | 14.35% |
| S-DCNet (cls) [47] | | ✓ | ✓ | | | | 58.3 | 95.0 | 13.94% |
| SS-DCNet (cls) | | ✓ | ✓ | ✓ | ✓ | ✓ | 57.8 | 100.8 | 13.92% |
| | | ✓ | ✓ | ✓ | ✓ | ✓ | 56.1 | 88.9 | 13.78% |

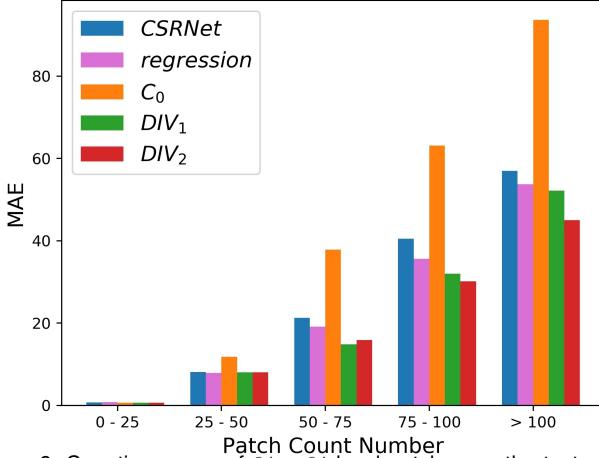


Figure 9. Counting errors of 64 × 64 local patches on the test set of ShanghaiTech Part_A [51]. CSRNNet [21] is a density map regression method which adopts VGG16 [38] as the feature extractor. *regression* denotes direct open-set local counts regression using VGG16 (*S_o* regression). *C₀*, *DIV₁* and *DIV₂* are count value predictions conditioned on F_0 , 1-stage division and 2-stage division of SS-DCNet (cls), respectively.

6.2.4 Choices of Loss Functions

Here we validate the effect of different loss functions used in SS-DCNet. Results are reported in Table 4. As analyzed in S-DCNet [47], L_C provides supervision to C_i s, and L_m implicitly supervises the division weights W_i s. With only L_C and L_m , S-DCNet (cls) can achieve good division results. However, S-DCNet (reg) cannot report competitive results as S-DCNet (cls) with L_R and L_m . After incorporating the upsampling loss L_{up} in the SS-DCNet (reg/cls), MAE reduces by 3.2/0.5 and rMAE reduces by 0.5%/0.02%. Such an improvement can be attributed to the replacement of the average upsampling in S-DCNet with learned upsampling in SS-DCNet. L_{div} provides explicit supervision for spatial division weight W_i s. One can see that, L_{div} can further improve the counting performance of SS-DCNet (reg/cls), and clear division results can be observed as shown in Fig. 10. Moreover, division consistency loss L_{eq} is also effective for SS-DCNet (reg), with 1.1 improvement in MAE and 1.11% in improvement in rMAE. Overall, SS-DCNet (reg/cls) shows a clear advantage over its previous version S-DCNet [47] with the help of additional supervision L_{up} , L_{div} and L_{eq} .

6.2.5 Spatial Divide-and-Conquer versus Spatial Attention

To highlight the difference between S-DC and spatial attention (SA), we remove the division decider, generate a 3-channel output conditioned on F_2 , then normalize it with softmax to obtain W_0^{att} , W_1^{att} and W_2^{att} . The final count is merged as $W_0^{att} * \text{upsample}(C_0) + W_1^{att} * \text{upsample}(C_1) + W_2^{att} * C_2$. In SHTech PartA, SA achieves 64.1 MAE and

Table 5
Configuration of SS-DCNet. \max denotes the maximum count of local patches in the training set. C_{\max} is the maximum count of the closed set in SS-DCNet. *Gaussian kernel* is used to generate density maps from dotted annotations. Specially, since UCF_CC_50 adopts 5-fold cross-validation, \max and C_{\max} are set adaptively for each fold.

| Dataset | C_{\max} | \max | Gaussian kernel |
|--------------------|------------|--------|----------------------|
| SH Part_A [51] | 22.0 | 148.5 | Geometry-Adaptive |
| UCF_CC_50 [14] | — | — | |
| UCF-QNRF [15] | 8.0 | 131.5 | |
| SH Part_B [51] | 7.0 | 83.0 | Fixed: $\sigma = 15$ |
| Trancos [13] | 5.0 | 24.5 | Fixed: $\sigma = 10$ |
| MTC [25] | 3.5 | 8.0 | Fixed: $\sigma = 8$ |
| Partition | | | Two-Linear |
| Type of C_{\max} | | | 95% quantile |

109.9 MSE, worse than SS-DCNet. As shown in the visualization of W_i^{att} in Fig. 10, we find SA only focuses on the highest resolution, and no effect of division is observed. Instead, SS-DCNet learns to divide local patches when local counts are greater than C_{\max} . In addition, SS-DCNet executes fusion recursively, while SA fuses the prediction in a single step.

6.3 Comparison with State of the Art Methods

According to the ablation study, the final configurations of SS-DCNet are summarized in Table 5.

6.3.1 The ShanghaiTech Dataset

The ShanghaiTech crowd counting dataset [51] includes two parts: Part_A and Part_B. Part_A has 300 images for training and 182 for testing. This part represents highly congested scenes. Part_B contains 716 images in relatively sparse scenes, where 400 images are used for training and 316 for testing. Quantitative results are listed in Table 6. The improvements of SS-DCNet are two-fold. First, with the explicit supervision of S-DC, SS-DCNet (cls) performs better than our previous S-DCNet. Second, our method outperforms the previous state-of-the-art PGCNet [49] in Part_A and competitive results (6.6 MAE) as SPANet [8] (6.5 MAE) in Part_B, respectively. These results suggest SS-DCNet is able to adapt to both sparse and crowded scenes.

6.3.2 The UCF_CC_50 Dataset

UCF_CC_50 [14] is a tiny crowd counting dataset with 50 images in extremely crowded scenes. The number of people within an images varies from 96 to 4633. We follow the 5-fold cross-validation as in [14]. Results are shown in Table 7. Our method surpasses S-DCNet and the previous best method, PaDNet [43], with 12.2% and 3.4% relative improvements in MAE, respectively.

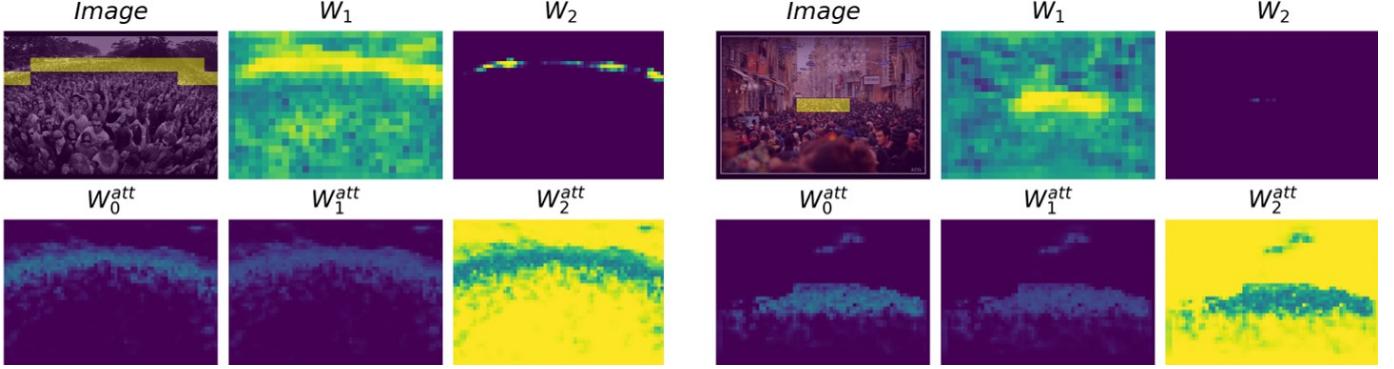


Figure 10. Visualization of W_i for SS-DCNet (cls) (top) and the attention baseline (bottom). The lighter the image is, the greater the values are. In the input image, count values greater than C_{max} are indicated by yellow regions.

Table 6
Performance on the test set of ShanghaiTech [51] dataset. The best performance is in boldface.

| Method | Part A | | | Part B | | |
|----------------|-------------|-------------|---------------|------------|------------|--------------|
| | MAE | MSE | rMAE | MAE | MSE | rMAE |
| IG-CNN [1] | 72.5 | 118.2 | — | 13.6 | 21.1 | — |
| DRSAN [24] | 69.3 | 96.4 | — | 11.1 | 18.2 | — |
| CSRNet [21] | 68.2 | 115.0 | 16.61% | 10.6 | 16.0 | 8.33% |
| SANet [3] | 67.0 | 104.5 | — | 8.4 | 13.6 | — |
| SPN [7] | 61.7 | 99.5 | — | 9.4 | 14.4 | — |
| BL [26] | 62.8 | 101.8 | 15.19% | 7.7 | 12.7 | 5.94% |
| PaDNet [43] | 59.2 | 98.1 | — | 8.1 | 12.2 | — |
| SPANet [8] | 59.4 | 92.5 | — | 6.5 | 9.9 | — |
| PGCNet [49] | 57.0 | 86.0 | — | 8.8 | 13.7 | — |
| S-DCNet [47] | 58.3 | 95.0 | 13.94% | 6.7 | 10.7 | 5.36% |
| SS-DCNet (reg) | 59.5 | 95.0 | 14.35% | 7.7 | 11.1 | 6.96% |
| SS-DCNet (cls) | 56.1 | 88.9 | 13.78% | 6.6 | 10.8 | 5.40% |

Table 8
Performance on the test set of UCF-QNRF [15] dataset. “r” denotes image resizing used in [26]. The best performance is in boldface.

| Method | r | MAE | MSE | rMAE |
|----------------|---|-------------|--------------|---------------|
| TEDnet [16] | x | 113 | 188 | — |
| CG-DRCN [40] | x | 112.2 | 176.3 | — |
| BL [26] | ✓ | 86.4 | 152.0 | 12.24% |
| PaDNet [43] | ✓ | 96.5 | 170.2 | — |
| CSRNet [21] | ✓ | 98.2 | 157.2 | 16.51% |
| S-DCNet [47] | x | 104.4 | 176.1 | 17.31% |
| S-DCNet [47] | ✓ | 97.7 | 167.6 | 14.58% |
| SS-DCNet (reg) | ✓ | 92.4 | 158.7 | 12.91% |
| SS-DCNet (cls) | ✓ | 81.9 | 143.8 | 12.64% |

6.3.3 The UCF-QNRF Dataset

UCF-QNRF [15] is a relatively large crowd counting dataset with 1535 high-resolution images and 1.25 million head annotations. There are 1201 training images and 334 test images. It contains extremely congested scenes where the maximum count of an image can reach 12865. Some images in the UCF-QNRF dataset are too large, with the longer side equals to 10000, to process the whole image. There are two ways to solve this problem: *i*) cropping the original image into 224×224 sub-images following [15]; *ii*) resizing the original image to make the longer side no larger than 1920 as in [48], [26], then 9 sub-images of $\frac{1}{4}$ resolution are cropped from the original image for data augmentation as described in Section 6.1.2. Results are reported in Table 8. We can make following observations:

- For S-DCNet, it works better with strategy *ii*) than *i*). This means that resizing is a better choice than cropping. We

Table 7
Performance on the test set of UCF_CC_50 [14] dataset. The best performance is in boldface.

| Method | MAE | MSE | rMAE |
|---------------------------|--------------|--------------|---------------|
| Idrees <i>et al.</i> [14] | 468.0 | 590.3 | — |
| Zhang <i>et al.</i> [50] | 467.0 | 498.5 | — |
| IG-CNN [1] | 291.4 | 349.4 | — |
| D-ConvNet [37] | 288.4 | 404.7 | — |
| CSRNet [21] | 266.1 | 397.5 | 30.22% |
| SANet [3] | 258.4 | 334.9 | — |
| SPANet [8] | 232.6 | 311.7 | — |
| DRSAN [24] | 219.2 | 250.2 | — |
| BL [26] | 213.8 | 310.5 | 20.46% |
| PaDNet [43] | 185.8 | 278.3 | — |
| S-DCNet [47] | 204.2 | 301.3 | 22.21% |
| SS-DCNet (reg) | 189.1 | 287.0 | 19.74% |
| SS-DCNet (cls) | 179.2 | 252.8 | 20.50% |

Table 9
Performance on the test set of TRANCOS [13] dataset. The best performance is in boldface.

| Method | GAME(0) | GAME(1) | GAME(2) | GAME(3) |
|----------------|-------------|-------------|-------------|-------------|
| CCNN [28] | 12.49 | 16.58 | 20.02 | 22.41 |
| Hydra-3s [28] | 10.99 | 13.75 | 16.69 | 19.32 |
| CSRNet [21] | 3.56 | 5.49 | 8.57 | 15.04 |
| SPN [7] | 3.35 | 4.94 | 6.47 | 9.22 |
| S-DCNet [47] | 2.92 | 4.29 | 5.54 | 7.05 |
| SS-DCNet (reg) | 2.73 | 3.82 | 5.05 | 6.72 |
| SS-DCNet (cls) | 2.42 | 3.30 | 4.55 | 6.17 |

Table 10
Performance on the test set of MTC [25] dataset. The best performance is in boldface.

| Method | MAE | MSE |
|-------------------------------|------------|------------|
| DensityReg [18] | 11.9 | 14.8 |
| CCNN [28] | 21.0 | 25.5 |
| TasselNet [25] | 6.6 | 9.6 |
| TasselNetv2 [†] [46] | 5.3 | 9.4 |
| CSRNet [21] | 5.4 | 7.9 |
| S-DCNet [47] | 5.6 | 9.1 |
| SS-DCNet (reg) | 4.0 | 6.9 |
| SS-DCNet (cls) | 3.9 | 6.6 |

think the reasons are two-fold. First, the receptive field of a CNN is limited, thus it cannot cover over-size images. Second, if cropping over-size images into 224×224 sub-images, the surrounding pixels of sub-images, termed ‘local visual context’, are invisible to the CNN. However, the local visual context can provide support information to distinguish overlapped objects as demonstrated in [46], and CNNs tend to perform poorly when local context is lost.

- With the explicit supervision of S-DC, SS-DCNet (cls) brings a significant improvement over S-DCNet by 15.8 in MAE, and SS-DCNet (reg) shows by 5.3 in MAE.
- SS-DCNet (reg) reports competitive results against the current state-of-the-art BL [26], while SS-DCNet (cls) outperforms BL by 6.8 in MAE and 11 in MSE.
- It is worth noting that, SS-DCNet only learns from a closed set with $C_{max} = 8.0$, which is only 6% of the maximum

count 131.5 according to Table 5. SS-DCNet, however, generalizes to large counts effectively and predicts accurate counts.

6.3.4 The TRANCOS Dataset

Aside from crowd counting, we also evaluate SS-DCNet on a vehicle counting dataset, TRANCOS [13], to demonstrate the generality of SS-DCNet. TRANCOS contains 1244 images of congested traffic scenes in various perspectives. It adopts the Grid Average Mean Absolute Error (GAME) [13] as the evaluation metric. $GAME(L)$ divides an image into $2^L \times 2^L$ non-overlapping sub-regions and accumulates of the MAE over sub-regions. Larger L implies more accurate local predictions. In particular, $GAME(0)$ downgrades to MAE . The GAME is defined by

$$GAME(L) = \frac{1}{N} \sum_{n=1}^N \left(\sum_{l=1}^{4^L} |C_{pre}^l - C_{gt}^l| \right), \quad (29)$$

where N denotes the number of images. C_{pre}^l and C_{gt}^l are the predicted and ground-truth count of the L -th sub-region, respectively. Results are listed in Table 9. SS-DCNet surpasses other methods under all $GAME(L)$ metrics, and particularly, delivers a 33.1% relative improvement than SPN [7] on $GAME(3)$. This suggests SS-DCNet not only achieves accurate global predictions but also behaves well in local regions.

6.3.5 The MTC Dataset

We further evaluate our method on a plant counting dataset, i.e., the MTC dataset [25]. The MTC dataset contains 361 high-resolution images of maize tassels collected from 2010 to 2015 in the wild field. In contrast to pedestrians or vehicles that have similar physical sizes, maize tassels are with heterogeneous physical sizes and are self-changing over time. We believe that this dataset is suitable for justifying the robustness of SS-DCNet to object-size variations. We follow the same setting as in [25] and report quantitative results in Table 10. Although the previous best method, TasselNetv2[†] [46], already exhibits accurate results, SS-DCNet still shows a substantial degree of improvement (26.4% on MAE and 29.8% on MSE).

Qualitative results are shown in Fig 11. We can observe that SS-DCNet produces accurate predictions for various objects from sparse to dense scenes.

6.4 Cross-Dataset Evaluation

We further conduct cross-dataset experiments on the ShanghaiTech [51] (A and B) and UCF-QNRF (QNRF) [15] datasets to show the generalization ability of SS-DCNet. Quantitative results are shown in Table 11. The ‘regression’ and ‘classification’ methods are baselines for SS-DCNet (reg) and S-DCNet (cls), respectively, which adopt the VGG16 [38] as the feature encoder and R-Counter/C-Counter in SS-DCNet but do not apply S-DC. We can make following observations:

- Consistent improvements in MAE are observed when comparing SS-DCNet (reg/cls) to its baselines. Especially in SS-DCNet (cls) vs. baseline cls, the MAE of $B \rightarrow QNRF$ shows a 40.89% relative improvement when S-DC is added.

- Two types of SS-DCNet report superior or at least competitive results than other state-of-the-art methods under all cross-dataset tasks, which suggest SS-DCNet has strong transferring ability.
- SS-DCNet (cls) transferred from the QNRF dataset reports even better results (61.8 MAE) than most state-of-the-arts methods (e.g., 68.2 MAE for CSRNet and 67.0 MAE for SANet) trained on the ShanghaiTech dataset.
- All methods trained on the ShanghaiTech [51] dataset report worse cross-dataset results than trained directly on the target datasets. By contrast, all methods trained on the QNRF [15] dataset exhibits at least competitive transferring results against state-of-the-art methods trained on the target dataset. This may be attributed to the fact that the ShanghaiTech dataset is too small, with only 300 training samples in the Part_A and 400 in the Part_B, to train a robust model, while the QNRF dataset provides sufficient training samples.

Overall, SS-DCNet demonstrates state-of-the-art results in all cross-dataset experiments. The good performance of SS-DCNet may be explained from its implicit transferring ability in the output space, which shares the same spirit with [44]. To justify this, we analyze the case of $QNRF \rightarrow A$ and visualize the distribution of count values with and without S-DC in Fig. 12. It can be observed that, the distribution of count values varies significantly between SHA and QNRF without S-DC, but the divergence of the distribution narrows down after count values are transformed into a closed set [0, 8]. We further compute the Jensen-Shannon divergence [29] J_s to quantify the divergence of the distribution, and find that $J_s = 0.0178$ between SHA and QNRF without S-DC and $J_s = 0.0133$ with S-DC. The smaller J_s is, the smaller divergence between two distributions shows. This means the divergence of the output (count value) space reduces after closed-set transformation. We believe this is the main reason why SS-DCNet reports remarkable performance on the task of $QNRF \rightarrow A$.

7 CONCLUSION

Counting is an open-set problem in theory, but only a finite closed set of training data can be observed in reality. This is particularly true because any dataset is always a sampling of the real world. Inspired by the decomposition property of counting, we have proposed to transform the open-set counting into a closed-set problem, and implement this transformation with the idea of S-DC. We propose supervised S-DC in a deep counting network, termed SS-DCNet. We provide a theoretical analysis showing why the transformation from the open set to closed set makes sense. Experiments on both synthetic data and real benchmark datasets show that, even given a closed training set, SS-DCNet can effectively generalize to open-set scenarios. Furthermore, SS-DCNet shows its good generalization ability via cross-dataset performance.

Many vision tasks are open-set by nature, depth estimation for example, while it is not immediately clear on how to transform them into a closed set like counting. It would be interesting to explore how to transform other vision tasks into a closed set setting.

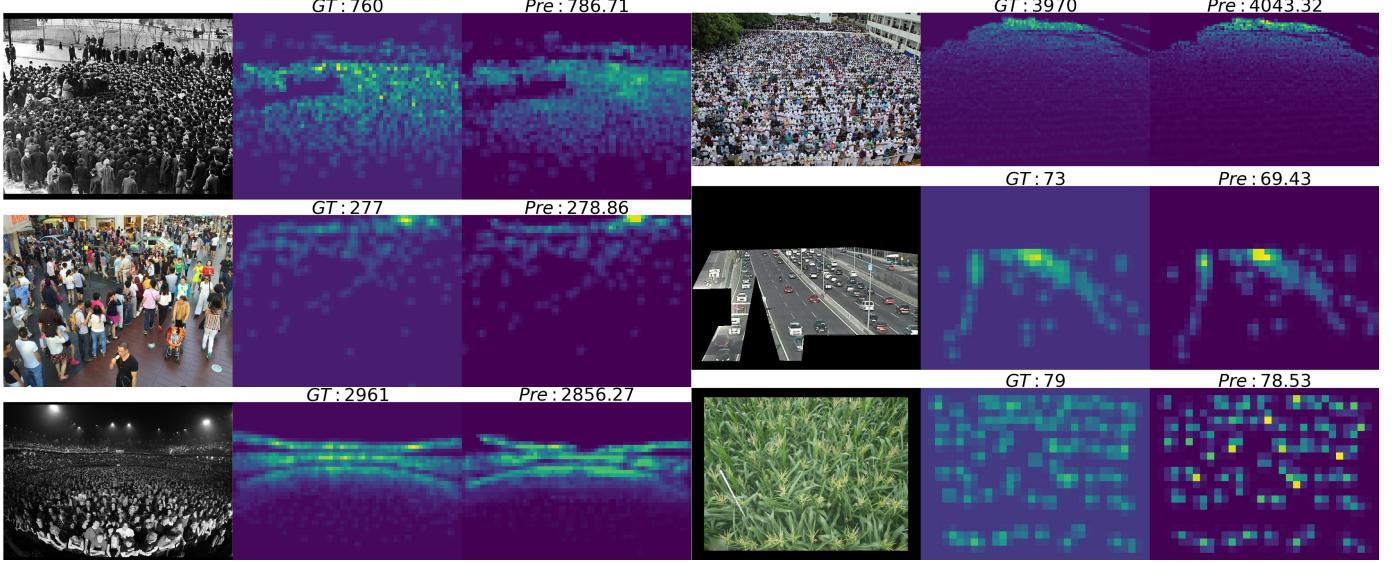


Figure 11. Visualization of count map for SS-DCNet on various datasets. From left to right, for each visualization, the original image, the ground-truth count map, and the inferred count map, respectively. From top to down, left to right, the visualization of ShanghaiTech Part_A, Part_B [51], UCF_CC_50 [14], UCF-QNRF [15], TRANCOS [13], and MTC [25], respectively. The original image is zero-padded to be divisible by 64.

Table 11

Cross-dataset performance (MAE/MSE/rMAE) on the ShanghaiTech (A and B) and UCF-QNRF (QNRF) Datasets. Best performance is in boldface.

| Method | A → B | A → QNRF | B → A | B → QNRF | QNRF → A | QNRF → B |
|----------------|-----------------------------|--------------------------------------|-------------------------------|---|--|-----------------------------------|
| D-ConvNet [37] | 49.1 / 99.2 / — | — / — / — | 140.4 / 226.1 / — | — / — / — | — / — / — | — / — / — |
| SPN+L2SM [48] | 21.2 / 38.7 / — | 227.2 / 405.2 / — | 126.8 / 203.9 / — | — / — / — | 73.4 / 119.4 / — | — / — / — |
| BL [26] | 16.3 / 30.3 / 12.24% | 141.6 / 252.4 / 23.41% | 137.0 / 228.9 / 30.32% | 208.9 / 414.4 / 25.97% | 69.8 / 123.8 / 14.85% | 15.3 / 26.5 / 10.99% |
| regression | 23.6 / 35.0 / 18.91% | 172.7 / 320.6 / 19.66% | 133.9 / 228.4 / 29.20% | 230.3 / 419.3 / 27.09% | 71.7 / 116.9 / 16.34% | 14.2 / 23.3 / 11.31% |
| classification | 21.4 / 36.6 / 16.49% | 173.7 / 323.1 / 21.64% | 179.4 / 313.4 / 31.60% | 281.7 / 512.3 / 28.49% | 134.7 / 259.5 / 22.81% | 21.9 / 47.8 / 14.24% |
| SS-DCNet (reg) | 22.9 / 35.2 / 19.55% | 160.6 / 299.3 / 19.27% | 137.1 / 235.8 / 30.77% | 222.1 / 399.5 / 27.98% | 69.0 / 115.8 / 15.53% | 12.1 / 20.8 / 9.20% |
| SS-DCNet (cls) | 21.2 / 39.5 / 15.89% | 151.8 / 270.4 / 18.36% | 130.0 / 209.6 / 27.82% | 166.5 / 281.8 / 23.14% | 61.8 / 102.8 / 13.79% | 11.8 / 21.8 / 7.98% |

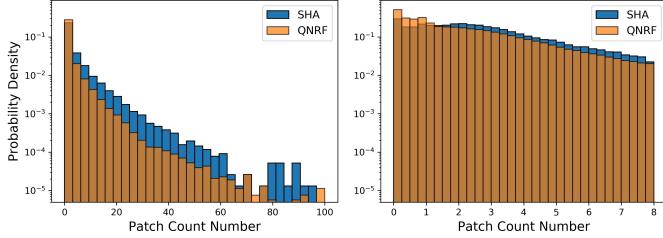
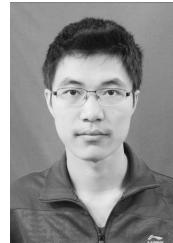


Figure 12. The probability density histograms of patch count numbers of SHA and QNRF. The area of each histogram adds up to 1. The left histogram shows the probability distribution without S-DC, and the right the probability distribution after S-DC, where count values are transformed into the closed-set [0, 8].

REFERENCES

- [1] Deepak Babu Sam, Neeraj N. Sajjan, R. Venkatesh Babu, and Mukundhan Srinivasan. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3618–3626, 2018.
- [2] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching convolutional neural network for crowd counting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5744–5752, 2017.
- [3] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *The European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.
- [4] Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R. Selvaraju, Dhruv Batra, and Devi Parikh. Counting everyday objects in everyday scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1135–1144, 2017.
- [5] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2467–2474, 2013.
- [6] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *Proc. British Machine Vision Conference (BMVC)*, 2012.
- [7] Xinya Chen, Yanrui Bin, Nong Sang, and Changxin Gao. Scale pyramid network for crowd counting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1941–1950, 2019.
- [8] ZhiQi Cheng, JunXiu Li, Qi Dai, Xiao Wu, and Alexander G Hauptmann. Learning spatial awareness to improve crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6152–6161, 2019.
- [9] Joseph Paul Cohen, Genevieve Boucher, Craig A. Glastonbury, Henry Z. Lo, and Yoshua Bengio. Count-ception: Counting by fully convolutional redundant counting. In *Proc. IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 18–26, 2017.
- [10] Stanislas Dehaene, Véronique Izard, Elizabeth Spelke, and Pierre Pica. Log or linear? distinct intuitions of the number scale in western and amazonian indigene cultures. *science*, 320(5880):1217–1220, 2008.
- [11] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2002–2011, 2018.
- [12] Ross Girshick. Fast R-CNN. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.
- [13] Ricardo Guerrerogómezolmedo, Beatriz Torrejiménez, Roberto Lópezastre, Saturnino Maldonadobascón, and Daniel Oñororubio. Extremely overlapping vehicle counting. In *Pattern Recognition and Image Analysis*, pages 423–431, 2015.
- [14] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2547–2554, 2013.

- [15] Haroon Idrees, Muhammad 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 *The European Conference on Computer Vision (ECCV)*, pages 532–546, 2018.
- [16] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao. Crowd counting and density estimation by trellis encoder-decoder networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6133–6142, 2019.
- [17] Mohamed A Khamsi and William A Kirk. *An introduction to metric spaces and fixed point theory*, volume 53. John Wiley & Sons, 2011.
- [18] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1324–1332, 2010.
- [19] Ruibo Li, Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, and Lingxiao Hang. Deep attention-based classification network for robust depth prediction. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2018.
- [20] Ruibo Li, Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, and Lingxiao Hang. Deep attention-based classification network for robust depth prediction. 2018.
- [21] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1091–1100, 2018.
- [22] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G. Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2018.
- [23] Liang Liu, Hao Lu, Haipeng Xiong, Ke Xian, Zhiguo Cao, and Chunhua Shen. Counting objects by blockwise classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.
- [24] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Lin Liang. Crowd counting using deep recurrent spatial-aware network. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.
- [25] Hao Lu, Zhiguo Cao, Yang Xiao, Bohan Zhuang, and Chunhua Shen. TasselNet: counting maize tassels in the wild via local counts regression network. *Plant Methods*, 13(1):79–95, 2017.
- [26] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6142–6151, 2019.
- [27] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4920–4928, 2016.
- [28] Daniel Oñoro-Rubio and Roberto J. López-Sastre. Towards perspective-free object counting with deep learning. In *The European Conference on Computer Vision (ECCV)*, pages 615–629, 2016.
- [29] Ferdinand Osterreicher and Igor Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653, 2003.
- [30] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 754–763, 2017.
- [31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.
- [32] Murray H Protter, B Charles Jr, et al. *A first course in real analysis*. Springer Science & Business Media, 2012.
- [33] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *The European Conference on Computer Vision (ECCV)*, pages 270–285, 2018.
- [34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241, 2015.
- [35] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2012.
- [36] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5245–5254, 2018.
- [37] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5382–5390, 2018.
- [38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.
- [39] Vishwanath A. Sindagi and Vishal M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1861–1870, 2017.
- [40] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1221–1231, 2019.
- [41] Tobias Stahl, Silvia L Pintea, and Jan C van Gemert. Divide and count: Generic object counting by image divisions. *IEEE Transactions on Image Processing*, 28(2):1035–1044, 2019.
- [42] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.
- [43] Y. Tian, Y. Lei, J. Zhang, and J. Z. Wang. Padnet: Pan-density crowd counting. *IEEE Transactions on Image Processing*, pages 1–1, 2019.
- [44] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [45] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.
- [46] Haipeng Xiong, Zhiguo Cao, Hao Lu, Simon Madec, Liang Liu, and Chunhua Shen. TasselNetv2: in-field counting of wheat spikes with context-augmented local regression networks. *Plant Methods*, 15:150–163, 2019.
- [47] Haipeng Xiong, Hao Lu, Chengxin Liu, Liu Liang, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8362–8371, 2019.
- [48] Chenfeng Xu, Kai Qiu, Jianlong Fu, Song Bai, Yongchao Xu, and Xiang Bai. Learn to scale: Generating multipolar normalized density maps for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8382–8390, 2019.
- [49] Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen, and Errui Ding. Perspective-guided convolution networks for crowd counting. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [50] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–841, 2015.
- [51] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–597, 2016.



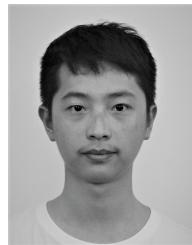
Haipeng Xiong received the B.S. degree from Huazhong University of science and Technology, Wuhan, China, in 2018. He is currently pursuing the M.S. degree with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China.

He currently researches object counting and its applications in agriculture. His research interests include math, machine learning and computer vision.



Hao Lu received the Ph.D. degree from Huazhong University of science and Technology, Wuhan, China, in 2018.

He is currently a Postdoctoral Fellow with the School of Computer Science, the University of Adelaide. His research interests include computer vision, image processing and machine learning. He has worked on topics including visual domain adaptation, fine-grained visual categorization, as well as miscellaneous computer vision applications in agriculture. His current interests are object counting and dense prediction problems.



Chengxin Liu received the B.S. degree from Huazhong University of science and Technology, Wuhan, China, in 2018. He is currently pursuing the M.S. degree with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. He currently researches object tracking.



Liang Liu received the B.S. degree from Huazhong University of science and Technology, Wuhan, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China.

His research interests include computer vision and machine learning, with particular emphasis on object counting and various computer vision applications in agriculture.

Chunhua Shen is a Professor of Computer Science, at the University of Adelaide, Australia.



Zhiguo Cao received the B.S. and M.S. degrees in communication and information system from the University of Electronic Science and Technology of China, Chengdu, China, and the Ph.D. degree in pattern recognition and intelligent system from Huazhong University of Science and Technology, Wuhan, China.

He is currently a Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. He has authored dozens of papers at international journals and conferences, which have been applied to automatic observation system for object recognition in video surveillance system, for crop growth in agriculture and for weather phenomenon in meteorology based on computer vision. His research interests spread across image understanding and analysis, depth information extraction and object detection.

Dr. Cao's projects have received provincial or ministerial level awards of Science and Technology Progress in China.