Semi-Supervised Crowd Counting via Self-Training on Surrogate Tasks

Yan Liu[1], Lingqiao Liu[2] ⋆, Peng Wang[3], Pingping Zhang[4], and Yinjie Lei[1] ⋆⋆

[1] College of Electronics and Information Engieering, Sichuan University
yanliu27@stu.scu.edu.cn, yinjie@scu.edu.cn
[2] School of Computer Science, The University of Adelaide
lingqiao.liu@adelaide.edu.au
[3] School of Computing and Information Technology, University of Wollongong
pengw@uow.edu.au
[4] School of Artificial Intelligence, Dalian University of Technology
jssxzhpp@mail.dlut.edu.cn

Abstract. Most existing crowd counting systems rely on the availability of the object location annotation which can be expensive to obtain. To reduce the annotation cost, one attractive solution is to leverage a large number of unlabeled images to build a crowd counting model in semi-supervised fashion. This paper tackles the semi-supervised crowd counting problem from the perspective of feature learning. Our key idea is to leverage the unlabeled images to train a generic feature extractor rather than the entire network of a crowd counter. The rationale of this design is that learning the feature extractor can be more reliable and robust towards the inevitable noisy supervision generated from the unlabeled data. Also, on top of a good feature extractor, it is possible to build a density map regressor with much fewer density map annotations. Specifically, we proposed a novel semi-supervised crowd counting method which is built upon two innovative components: (1) a set of inter-related binary segmentation tasks are derived from the original density map regression task as the surrogate prediction target; (2) the surrogate target predictors are learned from both labeled and unlabeled data by utilizing a proposed self-training scheme which fully exploits the underlying constraints of these binary segmentation tasks. Through experiments, we show that the proposed method is superior over the existing semi-supervised crowd counting method and other representative baselines.

Keywords: Crowd counting, surrogate tasks, self-training, semi-supervised learning

1 Introduction

Crowd counting is to estimate the number of people or objects from images or videos. Most existing methods formulate it as a density map regression problem

⋆ The first two authors have equal contribution.

⋆⋆ The corresponding author: Yinjie Lei (Email: yinijie@scu.edu.cn).

[1–5], and solve it by using the pixel-to-pixel prediction networks [6–8]. Once the density map is estimated, the total object count can be trivially calculated. To train such a density map regression model, most existing crowd counting methods rely on a substantial amount of labeled images with the object location annotation, e.g., marking a dot at the center of corresponding persons. The annotation process can be labor-intensive and time-consuming. For example, to annotate the ShanghaiTech [3] dataset, 330,165 dots must be placed on corresponding persons carefully.

To reduce the annotation cost, an attractive solution is to learn the crowd counter in a semi-supervised setting which assumes availability of a small amount of labeled images and a large amount of unlabeled images. This is a realistic assumption since unlabeled images are much easier or effortlessly to obtain than labeled images. Then the research problem is how to leverage the unlabeled image to help train the crowd counter for achieving a reasonable performance.

To solve this problem, we propose a novel semi-supervised learning algorithm to obtain a crowd counting model. One key of our model is to use the unlabeled data to learn a generic feature extractor of the crowd counter instead of the entire network as most traditional methods do. The underlying motivations are threefold: (1) It is challenging to construct a robust semi-supervised learning loss term from unlabeled data for regression output. In contrast, learning a feature extractor is more robust and reliable towards the inevitable noisy supervision generated from unlabeled data; (2) the feature extractor often plays a critical role in a prediction model. If we have a good feature extractor, it is possible to learn a density map regressor, i.e., crowd counter, require much less ground-truth density map annotations; (3) there are a range of methods for learning feature extractor, and features can be even learned from other tasks rather than density map regression (i.e., surrogate tasks in this paper).

Inspired by those motivations, we propose to learn the feature extractor through a set of surrogate tasks: predicting whether the density of a pixel is above multiple predefined thresholds. Essentially, those surrogate tasks are binary segmentation tasks and we build multiple segmentation predictors for each of them. Since those tasks are derived from the density map regression, we expect that through training with these surrogate tasks the network can learn good features to benefit the density map estimation. For labeled images, we have ground-truth segmentation derived from the ground-truth density map. For unlabeled images, the ground-truth segmentation are not available. However, the unlabeled images can still be leveraged through a semi-supervised segmentation algorithm. Also, we notice that the correct predictions for the surrogate tasks should hold certain inter-relationship, e.g., if the density of a pixel is predicted to be higher than a high threshold, it should also be predicted higher than a low threshold. Such inter-relationships could serve as additional cues for jointly training segmentation predictors under the semi-supervised learning setting. Inspired by that, we developed a novel self-training algorithm to incorporate these inter-relationships to generate reliable pseudo-labels for semi-supervised learning. By conducting extensive experiments, we demonstrate the superior performance of the proposed method. To sum up, our main contributions are:

- We approach the problem of semi-supervised crowd counting from a novel perspective of feature learning. By introducing the surrogate tasks, we cast the original problem into a set of semi-supervised segmentation problem.
- We develop a novel self-training method which fully takes advantage of the inter-relationship between multiple binary segmentation tasks.

2 Related Works

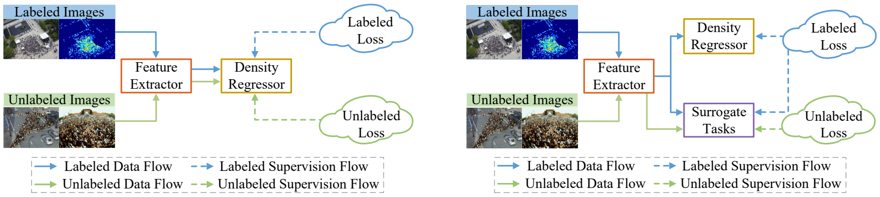
Traditional Crowd Counting Methods include detection-based and regression-based methods. The detection-based methods use head or body detectors to obtain the total count in an image [9–11]. However, in extremely congested scenes with occlusions, detection-based methods can not produce satisfying predictions.

Regression-based methods are proposed [12, 13] to tackle challenges in over-crowded scenes. In regression-based methods, feature extraction mechanisms [14, 15] such as Fourier Analysis and Random Forest regression are widely used. However, traditional methods can not predict total counts accurately as they overlook the spatial distribution information of crowds.

CNN-based Crowd Counting Methods learn a mapping function from the semantic features to density map instead of total count [1]. Convolutional Neural Network (CNN) shows great potential in computer vision tasks. CNN-based methods are used to predict density maps. Recently, the mainstream idea is to leverage deep neural networks for density regression [2–4, 16]. These methods construct multi-column structures to tackle scale variations. Then local or global contextual information is obtained for producing density maps.

Several works [5, 17] combine the VGG [18] structure with dilated convolution to assemble the semantic features for density regression. While other works [19–22] introduce attention mechanisms to handle several challenges, e.g. background noise and various resolutions. Meanwhile works [23–27] leverage the multi-task frameworks, i.e., detection, segmentation or localization, which provide more accurate location information for density regression. Besides, the self-attention mechanism [28, 29] and residual learning mechanism [30] are effective in regularizing the training of the feature extractor. Work [31] transforms the density value to the density level from close-set to open-set. Further, a Bayesian-based loss function [32] is proposed for density estimation. These above CNN-based methods require a large number of labeled images to train the crowd counter. However, annotating the crowd counting dataset is a time-consuming and labor-intensive work.

Semi-/Weakly/Un-Supervised Crowd Counting Methods attempt to reduce the annotation burden by using semi-/weakly/un-supervised settings. In the semi-supervised setting, work [33] collects large unlabeled images as extra training data and constructs a rank loss based on the estimated density maps. Also, work in [34] leverages the total count as a weak supervision signal for density estimation. Besides, an auto-encoder structure [35] is proposed for crowd counting in an almost unsupervised setting. Another method for reducing the annotation burden is to use synthetic images [36]. For example, the GAN-based [37] and domain adaption based [38] frameworks combine the synthetic images and realistic images to train the crowd counter. These methods are effective in reducing the annotation burden. However, they can not obtain satisfying crowd counting performance because the inevitable noisy supervision may mislead the density regressor.



(a) Traditional semi-supervised methods. (b) Proposed semi-supervised method.

Fig. 1. (a) Traditional semi-supervised methods use both labeled and unlabeled images to update the feature extractor and density regressor. (b) In the proposed method, the unlabeled images are only used for updating feature extractor.

3 Background: Crowd Counting as Density Estimation

Following the framework "learning to count" [1], crowd counting can be transformed into a density map regression problem. Once the density map is estimated, the total object count can be simply estimated by its summation, that is, $\hat{N} = \sum_{i,j} \hat{D}(i,j)$, where $\hat{D}(i,j)$ denotes the density value for pixel (i,j). The Mean Square Error (MSE) loss is commonly used in model training, that is,

$$\mathcal{L}_{MSE} = \sum_{(i,j)} |\hat{D}(i,j) - D(i,j)|^2, \tag{1}$$

where $\hat{D}$ is the estimated density map and D is the ground-truth density map.

4 Methodology

In this paper, we are interested in learning a crowd counter based on the semi-supervised setting. Formally, we assume that we have a set of labeled images $L = \{I_i^l, D_i\}$, where D_i is the ground-truth density map, and a set of unlabeled

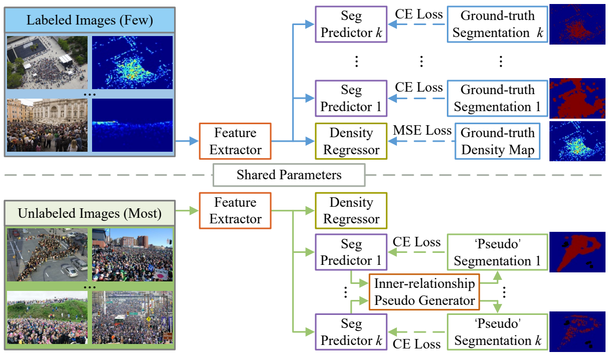


Fig. 2. The overview of our proposed method. We introduce a set of binary segmentation surrogate tasks. For labeled images, we construct loss terms on both original and surrogate tasks. For unlabeled images, we use the output of segmentation predictor and inter-relationship to generate "pseudo segmentation", which is shown in Figure. 3.

images $U = \{I_i^u\}$. Our task is to learn a crowd counter by using both the labeled images and unlabeled images. In our setting, the unlabeled set contains much more images than the labeled set for training a crowd counting model.

4.1 Using Unlabeled Data for Feature Learning

Generally speaking, a network can be divided into two parts, a feature extractor and a task-specific predictor. The former converts the raw images into feature maps while the latter further transforms them to the desired output, e.g., density map, in the context of crowd counting. Most existing semi-supervised learning methods [39, 33, 40, 41] learn those two parts simultaneously and seek to construct a loss term from unlabeled data applied to the entire network.

In contrast to the existing methods, we propose to learn the feature extractor and the task-specific predictor through different tasks and loss terms. In particular, in our method, the unlabeled data is only used for learning the feature extractor. This design is motivated by three considerations: (1) crowd counting is essentially a semi-supervised regression problem in our setting. Besides, it can be challenging to construct a robust semi-supervised regression loss term from unlabeled data (i.e., as most existing methods do). The noisy supervision generated from the loss term from unlabeled data may contaminate the task-specific predictor and lead to inferior performance. In our method, unlabeled data is only used to train the feature extractor as the noisy supervision will not directly affect the task-specific predictor; (2) feature extractor plays an important role in many fields like unsupervised feature learning [42, 43], semi-supervised feature learning [44, 45] and few-shot learning [46, 47]. Indeed, with a good feature extractor, it is possible to reduce the need of a large amount of labeled data in training. In the context of crowd counting, this implies that much less ground-truth density map annotations are needed if we can obtain a robust feature extractor via other means; (3) feature extractor can be learned in various ways. In this way, we will have more freedom in designing semi-supervised learning algorithms for feature learning. Specifically, we propose to derive surrogate tasks from the original density map regression problem, and use those tasks for training the feature extractor. The schematic overview of this idea is shown in Figure 1 (b). For labeled images, the target of surrogate task can be transformed from ground-truth annotation. For the unlabeled images, the ground-truth annotation becomes unavailable. However, the unlabeled images can still be leveraged to learn the surrogate task predictor and consequently the feature extractor in a semi-supervised learning manner. In the following sections, we first elaborate how to construct the surrogate loss and then describe the semi-supervised learning algorithm developed for the surrogate tasks.

4.2 Constructing Surrogate Tasks for Feature Learning

The surrogate task defined in this paper is to predict whether the density value of a pixel, $D(i,j)$, exceeds a given threshold. In other words, the prediction target of the surrogate task is defined as:

$$M(i,j) = \begin{cases} 1 & D(i,j) > \epsilon \\ 0 & D(i,j) <= \epsilon \end{cases}, \tag{2}$$

where (i,j) is the pixel coordinate, and ϵ is the predefined threshold. For labeled data, the ground-truth of D is known and thus M is known. For unlabeled data, no annotation of D is available and thus M is unknown. However, we can still use unlabeled data to construct loss term for indirectly supervising the prediction of M. Note that in this way, we essentially recast the original semi-supervised crowd counting problem into a semi-supervised segmentation problem since M only takes binary values.

In practice, we use multiple thresholds and generate multiple surrogate targets $\{M_k\}$ to consider the pixels with different density levels. To set these thresholds, we rank all non-zero density values from all the labeled images in ascending order and choose the thresholds as the value ranked at $r_k \times N$, where $r_k \in [0,1]$ $k = 1,..,c$, N is the total number of non-zero values and c indicates the number of surrogate tasks. Meanwhile, we create multiple segmentation predictor branches attached to the feature extractor. These surrogate tasks are parallel to the density map regressor, as shown in Figure 2.

4.3 Inter-Relationship-Aware Self-Training (IRAST) for Semi-supervised Training on Surrogate Tasks

To leverage the unlabeled data to train the surrogate task predictors and the feature extractor, a semi-supervised learning algorithm is needed. Self-training

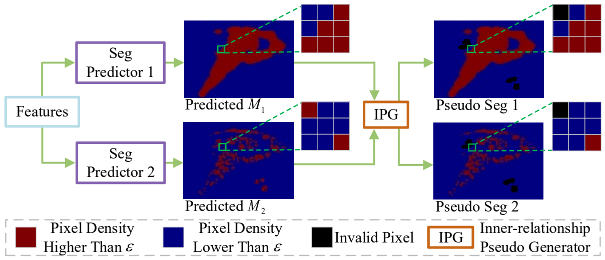

Seg Predictor 1
Features
Seg Predictor 2
Predicted M_1
Predicted M_2
IPG
Pseudo Seg 1
Pseudo Seg 2
Pixel Density Higher Than ε
Pixel Density Lower Than ε
Invalid Pixel
IPG
Inner-relationship Pseudo Generator

Fig. 3. The illustration of the inter-relationship between two segmentation predictors. We use a lower threshold ϵ_1 segmentation predictor to produce $\hat{M}_1$, and a higher threshold ϵ_2 segmentation predictor to produce $\hat{M}_2$. If a specific pixel in $\hat{M}_1$ is lower than ϵ_1, while in $\hat{M}_2$ is higher than ϵ_2, we can consider this pixel is invalid. The inter-relationship avoids such incorrect training signal flowing into the feature extractor.

is one of the most commonly used semi-supervised learning algorithms in segmentation tasks [48, 49]. It recursively generates pseudo-class-label for samples (pixels) with prediction confidence values higher than a given threshold $t_p = 0.9$. However, this straightforward solution largely ignores the underlying inter-relationship between multiple surrogate tasks. Recall that M takes binary values and $M(i, j) = 1$ if the density value of pixel (i, j) is greater than a given threshold. Suppose we have two segmentation results M_1 and M_2 estimated from two predictors corresponding to two thresholds ϵ_1 and ϵ_2 $(\epsilon_1 < \epsilon_2)$, then there will be a conflict if one predictor gives the prediction $\hat{M}_1(i, j) = 0$ while the other gives the prediction $\hat{M}_2(i, j) = 1$. This is because $\hat{M}_1(i, j) = 0$ indicates the density value of the pixel is less than ϵ_1, but $\hat{M}_2(i, j) = 1$ implies the density value of pixel is larger than ϵ_2 and consequently larger than ϵ_1 since $\epsilon_1 < \epsilon_2$.

This inter-relationship could essentially act as an error correcting mechanism to test if the prediction made by surrogate predictors are likely to be accurate. Thus in our method, we incorporate it into the framework of self-training as an additional criterion for pseudo-label generation besides the commonly used thresholding criterion. Formally, we define the following rule for generating a pseudo label at the k-th predictor. Without loss of generality, we assume there are c predictors, ranking from 1 to c according to the descent order of their corresponding thresholds, that is, $\epsilon_a > \epsilon_b$ if $a > b$. The formal rule of generating pseudo-label is shown in Algorithm 1.The generation of pseudo-labels is online.

In nutshell, a pseudo label is generated in the surrogate binary segmentation task if its prediction confidence value for one class ("1" or "0" in our case) is greater than t_p and its prediction is not conflict with predictions of other predictors. An example of this scheme is illustrated in Figure 3.

Discussion: The proposed method defines c binary segmentation tasks and one may wonder why not directly define a single c-way multi-class segmentation

Algorithm 1: Pseudo-label Generation Rule

```
Input: Number of surrogate tasks $c$. Given threshold $t_p$, Predicted confidence
       value (posterior probability) $P(\hat{M}_k = 1)$ $k = 1, \cdots, c$;
       $P(\hat{M}_k = 0) = 1 - P(\hat{M}_k = 1)$
Output: A set of pseudo-label set $\{\mathcal{S}_k\}$, one for each $k$: $\mathcal{S}_k = \{(i, j, s_{ij})\}$,
        where $s_{ij}$ is the generated pseudo-label for $(i, j)$.
for $k \in [1, c]$ do
    for each location $(i, j)$ do
        if $P(\hat{M}_k(i,j) = 1) > t_p$ and $P(\hat{M}_g(i,j) = 1) > t_p$ $\forall g < k$ then
            $\mathcal{S}_k \leftarrow \mathcal{S}_k \cup (i, j, 1)$
        end
        if $P(\hat{M}_k(i,j) = 0) > t_p$ and $P(\hat{M}_h(i,j) = 0) > t_p$ $\forall h > k$ then
            $\mathcal{S}_k \leftarrow \mathcal{S}_k \cup (i, j, 0)$
        end
    end
end
```

task. Then an standard multi-class self-training method can be used. We refer this method as **Multiple-class Segmentation Self-Training (MSST)**. Comparing with our approach, MSST has the following two disadvantages: (1) it does not have the "error correction" mechanism as described in the rule of generating pseudo label. The difference between MSST and IRAST is the standard one-vs-rest multi-class classification formulation and the error correcting output codes formulation [50]; (2) MSST may be overoptimistic towards the confidence score due to the softmax normalization of logits. Considering a three-way classification scenario for example, it is possible that the confidence for either class is low and the logits for all three classes are negative. But by chance, one class has relatively larger logits, say, $\{-100, -110$ and $-90\}$ for class 1, 2 and 3 respectively. After normalization, the posterior probability for the last class becomes near 1, and will exceed the threshold for generating pseudo labels. In contrast, the proposed IRAST does not have this issue since the confidence score will not be normalized across different classes (quantization level). We also conduct an ablation study in Section 6.3 to verify that MSST is inferior to IRAST.

5 Overall Training Process

In practice, we use the Stochastic Gradient Descent (SGD) to train the network[5]. For an labeled image, we construct supervised loss terms based on the density

[5] As the unlabeled set contains more images than the labeled set, we oversample labeled images to ensure the similar amount of labeled and unlabeled images occur in a single batch.

regression task and surrogate tasks, and the training loss is:

$$\mathcal{L}_L = \mathcal{L}_{MSE} + \lambda_1 \mathcal{L}_{SEG} = \sum_{(i,j)} \left(|\hat{D}(i,j) - D(i,j)|^2 + \lambda_1 \sum_{k=1}^{c} CE(M_k(i,j), \hat{M}_k(i,j)) \right),$$

where $CE()$ denotes the cross-entropy loss, $\hat{D}$ and $\hat{M}_k$ are the predicted density map and segmentation respectively; D and M_k are the ground-truth density map and segmentation respectively.

For an unlabeled image, we construct an unsupervised loss based on the surrogate tasks and use it to train the feature extractor:

$$\mathcal{L}_U = \lambda_2 \mathcal{L}_{SEG} = \lambda_2 \sum_{k=1}^{c} \sum_{(i,j,s_{ij}) \in \mathcal{S}_k} CE\left(\hat{M}_k(i,j)), s_{ij}\right), \qquad (3)$$

where the $\mathcal{S}_k = \{(i,j,s_{ij})\}$ denotes the set of generated pseudo labels at the k-th segmentation predictor. Please refer to Algorithm 1 for the generation of $\mathcal{S}_k$.

6 Experimental Results

We conduct extensive experiments on three popular crowd-counting datasets. The purpose is to verify if the proposed methods can achieve superior performance over other alternatives in a **semi-supervised learning setting** and understand the impact of various components of our method. Note that works that methods in a fully-supervised setting or a unsupervised setting are **not directly comparable** to ours.

6.1 Experimental Settings

Datasets. ShanghaiTech [3], UCF-QNRF [27] and WorldExpo'10 [2] are used throughout our experiments. We modify the setting of each dataset to suit the need of semi-supervised learning evaluation. Specifically, the original training dataset is divided into labeled and unlabeled sets. The details about such partition are given as follows.

ShanghaiTech [3]: The ShanghaiTech dataset consists of 1,198 images with 330,165 annotated persons, which is divided into two parts: Part_A and Part_B. Part_A is composed of 482 images with 244,167 annotated persons; the training set includes 300 images; the remaining 182 images are used for testing. Part_B consists of 716 images with 88,498 annotated persons. The size of the training set is 400, and the testing set contains 316 images. In Part_A, we randomly pick up 210 images to consist the unlabeled set, 90 images to consist the labeled set (60 images for validation). Also, In Part_B, we randomly pick up 280 images to consist the unlabeled set, 120 images to consist the labeled set (80 images for validation).

UCF-QNRF [27]: The UCF-QNRF dataset contains 1,535 high-resolution images with 1,251,642 annotated persons. The training set includes 1,201 images, and the testing set contains 334 images. We randomly pick up 721 images to consist the unlabeled set, 480 images to consist the labeled set (240 images for validation).

*World Expo'10 [2]:*The World Expo'10 dataset includes 3980 frames from Shanghai 2010 WorldExpo. The training set contains 3380 images, and the testing set consists of

600 frames. Besides, the Region of Interest (ROI) is available in each scene. Each frame and the corresponding annotated person should be masked with ROI before training. We randomly pick up 2433 images to consist the unlabeled set, 947 images to consist the labeled set (271 images for validation).

Compared Methods: We compare the proposed IRAST method against four methods: (1) Label data only (Label-only): only use the labeled dataset to train the network. This is the baseline of all semi-supervised crowd counting approaches. (2) Learning to Rank (L2R): a semi-supervised crowd counting method proposed in [33]. As the unlabeled images used in this paper are not released, we re-implement it with the same backbone and test setting as our method to ensure a fair comparison. (3) Unsupervised Data Augmentation (UDA): UDA [39] is one of the state-of-the-art semi-supervised learning methods. It encourages the network to generate similar predictions for an unlabeled image and its augmented version. This method was developed for image classification. We modify it by using the estimated density map as the network output. (4) Mean teacher (MT): Mean teacher [40] is a classic consistency-based semi-supervised learning approach. Similar as UDA, it was originally developed for the classification task and we apply it to the regression task by changing the network work output as the estimated density maps. (5) Interpolation Consistency Training (ICT): ICT [41] is a recently developed semi-supervised learning approach. It is based on the mixup data augmentation [51] but performed on unlabeled data. Again, we tailor it for the density map regression task by changing the output as the density map. More details about the implementation of the compared methods can be found in the supplementary material.

Implementation details: The feature extractor used in most of our experiment is based on the CSRNet [5]. We also conducted an ablation study in Section 6.3 to use Scale Pyramid Network (SPN) [17] as the feature extractor. Both CSRNet and SPN leverage VGG-16 [18] as the backbone. Also, three segmentation predictors are used by default unless specified. The thresholds for the corresponding surrogate tasks are selected as $\{0, 0.5N, 0.7N\}$ (please refer to Section 4.2 for the method of choosing thresholds). The segmentation predictors are attached to the 14-th layer of the CSRNet or the 13-th layer of SPN. The rest layers in those networks are viewed as the task specific predictor, i.e., the density map regressor. The segmentation predictors share the same network structure as the density map regressor. Please refer the supplementary material for the detailed structure of the network. In all experiments, we set the batch size as 1 and use Adam [52] as the optimizer. The learning rate is initially set to 1e-6 and halves per 30 epochs (120 epochs in total). Besides, we set t_p to 0.9 in experiments. Our implementation is based on PyTorch [53] and we will also release the code.

Evaluation metrics: Following the previous works [2, 3], the Mean Absolute Error (MAE) and Mean Squared Error (MSE) are adopted as the metrics to evaluate the performance of the compared crowd counting methods.

6.2 Datasets and Results

Evaluation on the ShanghaiTech Dataset: The experimental results on ShanghaiTech dataset are shown in Table 1. As seen, if we only use the labeled image, the network can only attain an MAE of 98.3 on Part_A and 15.8 on Part_B. In general, using a semi-supervised learning approach brings improvement. The L2R [33] shows an improvement around 8 people in the MAE of Part A but almost no improvement for Part B. Semi-supervised learning approaches modified from the classification task (UDA, MT, ICT) also lead to improved performance over Label-only on Part A. However, the improvement is not as large as L2R. Our approach, IRAST, clearly demonstrates the best performance. It leads to 11.4 MAE improvement over the Label-only on Part A.

Table 1. The comparison on the ShanghaiTech dataset. The best results are in bold font.

	Method	Part_A		Part_B	
		MAE	MSE	MAE	MSE
	Label-only	98.3	159.2	15.8	25.0
Semi	L2R [33]	90.3	153.5	15.6	24.4
	UDA [39]	93.8	157.2	15.7	24.1
	MT [40]	94.5	156.1	15.6	24.5
	ICT [41]	92.5	156.8	15.4	23.8
	IRAST	**86.9**	**148.9**	**14.7**	**22.9**
	(Fully) CSRNet [5]	68.2	115.0	10.6	16.0

Table 2. The comparison on the UCF-QNRF dataset. The best results are in bold font.

	Method	UCF-QNRF	
		MAE	MSE
	Label-only	147.7	253.1
Semi	L2R [33]	148.9	249.8
	UDA [39]	144.7	255.9
	MT [40]	145.5	250.3
	ICT [41]	144.9	250.0
	IRAST	**135.6**	**233.4**
	(Fully) CSRNet [5]	119.2	211.4

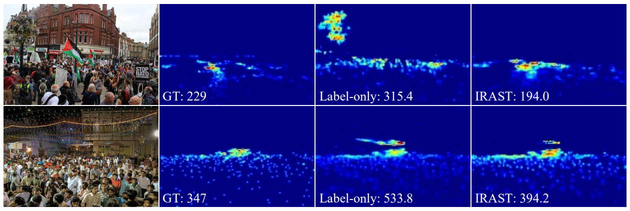


Fig. 4. A comparison of predicted density maps on the UCF-QNRF dataset.

Evaluation on the UCF-QNRF Dataset: The advantage of the proposed method is also well demonstrated on UCF-QNRF dataset, shown in Table 2. Again, the proposed method achieves the overall best performance, and exceeds the Label-only by around 12 MAE. The other semi-supervised learning approach does not work well on this dataset. In particular, L2R even achieves worse performance than the Label-only. This on the other hand clearly demonstrates the robustness of our approach. Also, from the results in both ShanghaiTech and UCF-QNRF, we can see that directly employing the semi-supervised learning approaches which were originally developed for classification may not achieve satisfying performance. It remains challenging for developing the semi-supervised crowd counting algorithm.

Evaluation on the World Expo'10 Dataset: The results are shown in Table 3. As seen, IRAST again achieves the best MAE in 2 scenes and delivers the best MAE over other methods. The other semi-supervised learning methods achieve comparable performance and their performance gain over the Label-only is not significant.

Table 3. The performance comparison in terms of MAE on the WorldExpo'10 dataset. The best results are in bold font.

	Method	Sce.1	Sce.2	Sce.3	Sce.4	Sce.5	Avg.
	Label-only	2.4	16.9	9.7	41.3	3.1	14.7
Semi	L2R [33]	2.4	20.9	9.8	31.9	4.4	13.9
	UDA [39]	**1.9**	20.3	10.9	34.5	3.6	14.2
	MT [40]	2.6	24.8	9.4	30.3	3.3	14.1
	ICT [41]	2.3	17.8	**8.3**	43.5	**2.8**	14.9
	IRAST	2.2	**12.3**	9.2	**27.8**	4.1	**11.1**
	(Fully) CSRNet [5]	2.9	11.5	8.6	16.6	3.4	8.6

6.3 Ablation Study

To understand the importance of various components in our algorithm, we conduct a serials of ablation studies.

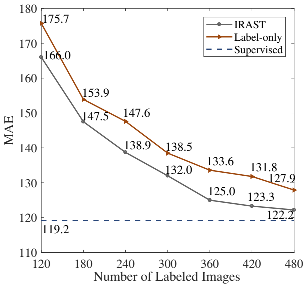


Fig. 5. The impact of the number of labeled images. Evaluated in terms of MAE on the UCF-QNRF dataset.

Varying the Number of Labeled Images: We first examine the performance gain over the Label-only under different amount of labeled images. We conduct experiments on the UCF-QNRF dataset. We vary the number of labeled image from 120 to 480 while fixing the amount of unlabeled images to be 481. The performance curves of the Label-only and IRAST are depicted in Figure 5. As seen, IRAST achieves consistent performance gain over the Label-only, which is an evidence of the robustness of our method. Also, we can see that with IRAST, using 480 images can almost achieve comparable performance than the performance of a fully-supervised model which needs 961 training images.

IRAST on Labeled set: The proposed method constructs an additional training task and one may suspect the good performance is benefited from the multi-task learning. To investigate this hypothesis, we also conduct an ablation study by learning the crowd counter on the labeled set only, but with both density map regression task and surrogate tasks. The results are shown in Table 4. As seen, using multiple-surrogate tasks for the labeled set does improve the performance to some extent, but still has a significant performance gap with the proposed method. This result clearly validates that our method can not be simply understood as a multi-task learning approach.

Other Alternative Surrogate Task: One alternative method is to use a multi-class segmentation predictor to train the feature extractor, namely MSST mentioned in Section 4.3. To compare MSST and IRAST, we conduct experiments on the ShanghaiTech Part_A and UCF-QNRF dataset. The results are shown in Table 5. As seen, MSST can achieve a better performance than Label-only method, which demonstrates the effectiveness of using a surrogate task for feature learning. However, MSST obtains a worse crowd counting performance than IRAST. Recall that MSST lacks an error correction mechanism to generate pseud-label, the superior performance of IRAST over MSST provides evidence to support the merit of our multiple surrogate binary-segmentation task modelling.

Table 4. Impact of the unlabeled images in the process of feature learning. Evaluated on the ShanghaiTech Part_A and UCF-QNRF dataset. The best results are in bold font.

Method	Part_A		UCF-QNRF	
	MAE	MSE	MAE	MSE
Label-only	98.3	159.2	147.7	253.1
IRAST on label	94.1	151.6	140.8	245.4
IRAST	**86.9**	**148.9**	**135.6**	**233.4**

Table 5. Comparison of IRAST and MSST. Evaluated on the ShanghaiTech Part_A and UCF-QNRF dataset. The best results are in bold font.

Method	Part_A		UCF-QNRF	
	MAE	MSE	MAE	MSE
Label-only	98.3	159.2	147.7	253.1
MSST	91.5	155.2	140.0	233.7
IRAST	**86.9**	**148.9**	**135.6**	**233.4**

Table 6. Impact of the inter-relationship. Evaluated on the ShanghaiTech Part_A and UCF-QNRF dataset. The best results are in bold font.

Method	Part_A		UCF-QNRF	
	MAE	MSE	MAE	MSE
Label-only	98.3	159.2	147.7	253.1
IRAST w/o IR	93.5	155.5	139.8	240.3
IRAST	**86.9**	**148.9**	**135.6**	**233.4**

Table 7. Impact of the changing hyper-parameter t_p. Evaluated on the ShanghaiTech Part_A and UCF-QNRF dataset. The best results are in bold font.

Method	Part_A		UCF-QNRF	
	MAE	MSE	MAE	MSE
Label-only	98.3	159.2	147.7	253.1
$t_p = 0.6$	88.4	152.3	137.2	234.9
$t_p = 0.9$	**86.9**	**148.9**	**135.6**	**233.4**

The Importance of Considering the Inter-Relationship: In IRAST, we leverage the Inter-Relationship (IR) between surrogate tasks to generate pseudo-labels. To verify the importance of this consideration, we conduct an ablation study by removing the inter-Relationship constraint for pseudo-label generation. The results are shown in Table 6. As seen, a decrease in performance is observed when the Inter-Relationship is not considered. This observation suggests that the Inter-Relationship awareness is essential to the proposed IRAST method.

The Impact of Changing the Prediction Confidence Threshold: We set hyper-parameter t_p to 0.9 in the previous experiments. To investigate the impact of t_p, we conduct experiments on ShanghaiTech part_A and UCF-QNRF dataset. The results are shown in Table 7. The results demonstrate setting diverse t_p does not impact crowd counting performance significantly. The crowd counting performance are comparable to our current results. It means the proposed IRAST method is robust.

Table 8. Impact of the feature extractor. Evaluated on the ShanghaiTech Part_A and UCF-QNRF dataset. The best results are in bold font.

Method	Part_A		UCF-QNRF	
	MAE	MSE	MAE	MSE
Label-only (CSRNet)	98.3	159.2	147.7	253.1
IRAST (CSRNet)	86.9	148.9	135.6	233.4
Label-only (SPN)	88.5	152.6	138.0	244.5
IRAST (SPN)	**83.9**	**140.1**	**128.4**	**225.3**

Table 9. Impact of the varing number of surrogate tasks. The best results are in bold font.

Tasks	Part_A		UCF-QNRF	
	MAE	MSE	MAE	MSE
1	89.8	149.8	142.8	236.5
2	88.9	149.6	139.1	237.8
3	**86.9**	**148.9**	**135.6**	**233.4**
4	90.1	150.2	137.5	236.8
5	90.3	150.9	137.8	234.4

Change of the Feature Extractor: So far, we conduct our experiment with the CSRNet [5] feature extractor. It is unclear if performance gain can still be achieved with other feature extractors. To investigate this, we conduct an experiment that uses SPN [17] as the feature extractor on the ShanghaiTech Part_A and UCF-QNRF dataset. Results are shown in Table 8. We can see that the significant performance gain can still be achieved. Also, we observe an improved overall performance by using SPN. This suggests that the advances in network architecture design for crowd counting can be readily incorporated into our method.

The Effect of Varying the Number of Surrogates Tasks: Finally, we test the impact of choosing the number of surrogate tasks. We incrementally adding more thresholds by following the threshold sequence $\{0, 0.5N, 0.7N, 0.8N, 0.9N\}$, e.g., $\{0, 0.5N, 0.7N\}$ is used for the three-task setting while $\{0, 0.5N, 0.7N, 0.8N\}$ is used for the four-task setting. The results are shown in Table 9. The results demonstrate setting three surrogate tasks for feature learning can achieve the best crowd counting performance. To have a finer grained partition of density value does not necessarily lead to improved performance.

7 Conclusions

In this paper, we proposed a semi-supervised crowd counting algorithm by creating a set of surrogate tasks for learning the feature extractor. A novel self-training strategy that can leverage the inter-relationship of different surrogate tasks is developed. Through extensive experiments, it is clear that the proposed method enjoys superior performance over other semi-supervised crowd counter learning approaches.

Acknowledgement

This work was supported by the Key Research and Development Program of Sichuan Province (2019YFG0409). Lingqiao Liu was in part supported by ARC DECRA Fellowship DE170101259.

References

1. Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1324–1332, 2010.
2. Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–841, 2015.
3. Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–597, 2016.
4. Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching convolutional neural network for crowd counting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5744–5752, 2017.
5. Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1091–1100, 2018.
6. Kai Kang and Xiaogang Wang. Fully convolutional neural networks for crowd segmentation. *arXiv preprint arXiv:1411.4464*, 2014.
7. Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4353–4361, 2017.
8. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.
9. Venkatesh Bala Subburaman, Adrien Descamps, and Cyril Carincotte. Counting people in the crowd using a generic head detector. In *IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pages 470–475, 2012.
10. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005.
11. Paul Viola and Michael J Jones. Robust real-time face detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, 2004.
12. Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *British Machine Vision Conference (BMVC)*, pages 21.1–21.11, 2012.
13. Antoni B Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. In *IEEE International Conference on Computer Vision (ICCV)*, pages 545–551, 2009.
14. Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2547–2554, 2013.
15. Luca Fiaschi, Ullrich Köthe, Rahul Nair, and Fred A Hamprecht. Learning to count with regression forest and structured labels. In *International Conference on Pattern Recognition (ICPR)*, pages 2685–2688, 2012.
16. Vishwanath A. Sindagi and Vishal M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1861–1870, 2017.

17. Xinya Chen, Yanrui Bin, Nong Sang, and Changxin Gao. Scale pyramid network for crowd counting. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1941–1950, 2019.
18. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.
19. 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6133–6142, 2019.
20. Muming Zhao, Jian Zhang, Chongyang Zhang, and Wenjun Zhang. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12736–12745, 2019.
21. Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *European Conference on Computer Vision (ECCV)*, pages 270–285, 2018.
22. Liang Zhu, Zhijian Zhao, Chao Lu, Yining Lin, Peng Yao, and Tangren Yao. Dual path multi-scale fusion networks with attention for crowd counting. *arXiv preprint arXiv:1902.01115*, 2019.
23. Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G. Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9175–9184, 2018.
24. Dongze Lian, Jing Li, Jia Zheng, Weixin Luo, and Shenghua Gao. Density map regression guided detection network for rgb-d crowd counting and localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1821–1830, 2019.
25. Shenqin Jiang, Xiaobo Lu, Yinjie Lei, and Lingqiao Liu. Mask-aware networks for crowd counting. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2019.
26. Varun Kannadi Valloli and Kinal Mehta. W-net: Reinforced u-net for density map estimation. *arXiv preprint arXiv:1903.11249*, 2019.
27. Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *European Conference on Computer Vision (ECCV)*, pages 532–546, 2018.
28. Anran Zhang, Jiayi Shen, Zehao Xiao, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Relational attention network for crowd counting. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6788–6797, 2019.
29. Jia Wan and Antoni Chan. Adaptive density map generation for crowd counting. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1130–1139, 2019.
30. Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1221–1231, 2019.
31. 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 *IEEE International Conference on Computer Vision (ICCV)*, pages 8362–8371, 2019.
32. Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6142–6151, 2019.

33. Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7661–7669, 2018.
34. Matthias von Borstel, Melih Kandemir, Philip Schmidt, Madhavi K Rao, Kumar Rajamani, and Fred A Hamprecht. Gaussian process density counting from weak supervision. In *European Conference on Computer Vision (ECCV)*, pages 365–380, 2016.
35. Deepak Babu Sam, Neeraj N Sajjan, Himanshu Maurya, and R Venkatesh Babu. Almost unsupervised learning for dense crowd counting. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 8868–8875, 2019.
36. Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8198–8207, 2019.
37. Junyu Gao, Qi Wang, and Yuan Yuan. Feature-aware adaptation and structured density alignment for crowd counting in video surveillance. *arXiv preprint arXiv:1912.03672*, 2019.
38. Junyu Gao, Tao Han, Qi Wang, and Yuan Yuan. Domain-adaptive crowd counting via inter-domain features segregation and gaussian-prior reconstruction. *arXiv preprint arXiv:1912.03677*, 2019.
39. Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*, 2019.
40. Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1195–1204, 2017.
41. Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3635–3641, 2019.
42. Fan Zhang, Bo Du, and Liangpei Zhang. Saliency-guided unsupervised feature learning for scene classification. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 53(4):2175–2184, 2014.
43. Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 766–774, 2014.
44. Yang Yang, Guang Shu, and Mubarak Shah. Semi-supervised learning of feature hierarchies for object detection in a video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1650–1657, 2013.
45. Yanhua Cheng, Xin Zhao, Kaiqi Huang, and Tieniu Tan. Semi-supervised learning for rgb-d object recognition. In *International Conference on Pattern Recognition (ICPR)*, pages 2377–2382, 2014.
46. Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *IEEE International Conference on Computer Vision (ICCV)*, pages 8059–8068, 2019.
47. Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10, 2019.
48. Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *IEEE Computer*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 966–973, 2010.

49. Stefan Karnyaczki and Christian Desrosiers. A sparse coding method for semi-supervised segmentation with multi-class histogram constraints. In *IEEE International Conference on Image Processing (ICIP)*, pages 3215–3219, 2015.
50. Thomas G Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research (JAIR)*, 2:263–286, 1994.
51. Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.
52. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, pages 1–13, 2014.
53. 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 (NIPS)*, pages 8024–8035, 2019.