Spatial Uncertainty-Aware Semi-Supervised Crowd Counting

Anonymous ICCV submission

Paper ID 1834

Abstract

Semi-supervised approaches for crowd counting attract attention, as the fully supervised paradigm is expensive and laborious due to its request for a large number of images of dense crowd scenarios and their annotations. This paper proposes a spatial uncertainty-aware semi-supervised approach via regularized surrogate task (binary segmentation) for crowd counting problems. Different from existing semi-supervised learning-based crowd counting methods, to exploit the unlabeled data, our proposed spatial uncertainty-aware teacher-student framework focuses on high confident regions' information while addressing the noisy supervision from the unlabeled data in an end-to-end manner. Specifically, we estimate the spatial uncertainty maps from the teacher model's surrogate task to guide the feature learning of the main task (density regression) and the surrogate task of the student model at the same time. Besides, we introduce a simple yet effective differential transformation layer to enforce the inherent spatial consistency regularization between the main task and the surrogate task in the student model, which helps the surrogate task to yield more reliable predictions and generates high-quality uncertainty maps. Thus, our model can also address the task-level perturbation problems that occur spatial inconsistency between the primary and surrogate tasks in the student model. Experimental results on four challenging crowd counting datasets demonstrate that our method achieves superior performance to the state-of-the-art semi-supervised methods.

1. Introduction

The task of crowd counting in computer vision is to infer the number of people in images or videos. There is an ever-increasing demand for automated crowd counting techniques in various applications such as public safety, security alerts, transport management *etc.*.

With the help of Convolutional Neural Network (CNN)'s feature learning ability, current state-of-the-art methods [1, 51, 47, 57, 39, 36] gained excellent counting performance by regressing the corresponding density maps of the input images, where the summed value in a density map gives the total count numbers. To train a robust and accurate crowd counting estimator, most of the existing methods [18, 34, 22, 17, 7, 33] relied on substantial labeled images, where head centres must be annotated for training. However, the annotation process can be labour-intensive and time-consuming. For example, JHU-Crowd [42] dataset contains labels of 1.51 millions people whilst NWPU-Crowd [48] dataset contains annotations of 2.13 millions people, which takes 3000 human hours in total. Hence, reducing annotation efforts while maintaining good counting performance is our goal in this paper. More specifically, we study the counting estimator in a semi-supervised manner where limited labeled data is used; on the other hand, the unlabeled data is leveraged to improve our model's robustness and performance.

Previous semi-supervised crowd counting methods tend to minimize the expensive label work through active learning [60, 20], synthetic images [49, 50], or pseudo-ground truth [40, 21]. However, they did not consider the unlabeled data or synthetic data's intrinsic noisy supervision due to the inherent data uncertainties [27]. Uncertainty estimation has been explored in other computer vision tasks, such as segmentation [10, 53, 2] or detection [56, 5], *etc.*. There are two significant types of uncertainty [9]: epistemic uncertainty, which accounts for the uncertainty in the model parameters and can be addressed when given enough data; aleatoric uncertainty corresponds to inevitable noisy perturbation existing in the data itself. Solving the aleatoric uncertainty is a crucial problem since crowd images contain inherent noises such as complex backgrounds, massive occlusions and illumination variations *etc.*. Few recent approaches [27, 28] have considered the uncertainty quantification in the crowd counting task in a fully-supervised manner. They adopted [9] to estimate the mean and variance of the assumed Gaussian distribution of the density map, where the variance is served as a measure of uncertainty.

In this work, we exploit the aleatoric uncertainty in a semi-supervised manner to alleviate the noisy supervision in uncertain spatial regions due to the complex back-

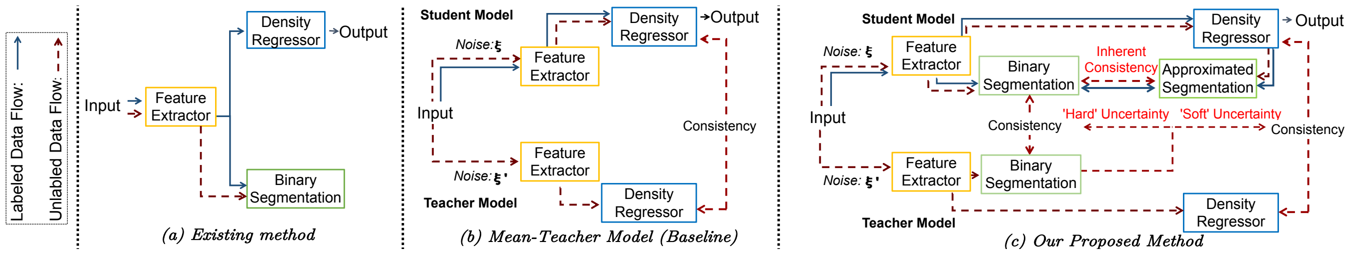

Figure 1. Overview of the very recent work [21], baseline model [44], and our proposed method. (a): [21] utilized surrogate task (binary segmentation) to boost the feature extractor with labeled and unlabeled data so as to enhance the performance of the density regressor. (b): Mean-Teacher [44] is a commonly used semi-supervised framework through exploiting the consistency learning on the student and the teacher models' outputs under different input-level noise perturbations (ξ, ξ') and model-level noise perturbation (Dropout [43] of the student and teacher models). We refer to it as the baseline model in this paper. (c): Our Mean-Teacher based semi-supervised framework. Note that we only input the unlabeled data into the teacher model because this work aims to explore the unlabeled data's uncertainty. The estimated 'hard' and 'soft' spatial uncertainty maps aim to assist the consistency learning (upon binary segmentation and density regression) between the student and teacher models; one can alleviate the unlabeled data's inevitable noisy supervision. The student model's binary segmentation is regularized by the inherent consistency regularization with approximated segmentation to address the spatial predictions' perturbation issues between binary segmentation and density regression tasks in the student model.

grounds and massive occlusions challenges from the unlabeled crowd images [27]. Previous crowd counting methods [59, 35, 4] prove that the spatial region information from the binary segmentation task is essential to tell the crowd and background locations, which will help the density map regressor to focus on the region of interest and improve the counting performance. In our work, the binary segmentation provides spatial information and serves as a surrogate task to estimate the uncertain spatial regions ($e.g.$ uncertain crowd locations). With the estimated spatial uncertainty, we assist the unsupervised consistency learning (upon binary segmentation and density regression) between the student model and the teacher model based on the Mean-Teacher [44] semi-supervised learning framework. Fig 1 (b & c) shows the overview structure of our method and the re-implemented Mean-Teacher framework [44] for the crowd counting task. Note that, in our work, the student model and the teacher model share a similar structure (Feature extractor, binary segmentation module, density regressor). We update the teacher model's parameters as an exponential moving average (EMA) of the student model's parameters. Because ensembling the student model's predictions at different training steps can enhance the performance of the teacher model's predictions [11]; in which case, the teacher model can generate 'targets' for the student model to learn from. However, as mentioned above, those 'targets' contain spatial-wise uncertainty; thus, we purify the 'targets' with the estimated 'hard' and 'soft' uncertainty map during training.

Apart from the aforementioned novel components, we also study how to learn an excellent surrogate task (binary segmentation) predictor to produce reliable and consistent spatial uncertainty that the main task (density regression) has in the student model. Note that, followed by [44], only student model is used for the inference process. Specifically, we introduce a simple yet effective differentiable transformation layer to approximate the binary segmentation maps from the density map predictions of the unlabeled input in the student model. We then employ an unsupervised inherent consistency loss between the predicted segmentation maps and the approximated segmentation maps to guarantee the consistent spatial feature learning between two different tasks in the student model. The underlying motivations are twofold: (1) the surrogate and the main task may introduce an inherent prediction perturbation on spatial regions due to the domain gap of feature learning from multi-tasks [25]. Our ablation experiment results prove that this perturbation will lead to noisy supervision upon two tasks, thus reducing the performance. (2) The proposed transformation layer itself is simple. However, it brings several benefits with the inherent consistency loss. For example, the estimated uncertainty from a regularized surrogate task can indicate more reasonable and consistent spatial uncertain regions that the main task has, which further enhances the consistency between the surrogate and the main task. In other words, with the proposed transformation layer, the estimated uncertainty and consistency regularization can benefit from each other to advance the counting performance. Our experiment results demonstrate that the proposed consistency regularization mechanism can boost the model's performance in both supervised and semi-supervised manner.

In summary, this work makes the following contributions: (1) We propose a surrogate task to estimate the uncertain spatial regions from the unlabeled data under the semi-supervised Teacher-Student framework, which can alleviate the inevitable noisy supervision from the unlabeled data. (2) We propose a differentiable transformation layer that enables the inherent spatial consistency regularization between the surrogate task (binary segmentation) and the main task (density regression) in the student model, which can enhance the model to estimate high-quality uncertainty maps from the unlabeled data, thus improve our model's counting performance. (3) We conduct extensive experiments on four well-known challenging counting benchmarks. Quantitative results demonstrate that our methods outperform existing semi-supervised crowd counting methods. Besides, with less than half of the labeled data, our method can achieve comparable performance with the fully-supervised state-of-the-art methods.

2. Related Works

2.1. Supervised Density-based Crowd Counting

Recently, fully-supervised density map regression-based counting methods with CNN achieved good performance. Approaches like [3, 58, 55] proposed a multi-column network to regress the density map in terms of combining local and global features to tackle the scale variation challenges. Other works [26, 8, 54] employed visual attention mechanisms to address other issues, such as background noise in crowded cluster scenarios and various density levels from scale variations. Apart from single-task learning, recent works introduced auxiliary task learning frameworks, i.e. classification [35, 38], localization [16, 29, 15, 14, 24], or segmentation [59, 35, 4], which attains additional spatial and semantic information supplement from the joint learning auxiliary tasks. The above methods focus on improving the counting performance in a fully-supervised paradigm. However, annotating the crowd counting dataset is labour-intensive and time-consuming work. In this paper, we made efforts on minimizing the expensive labelling work in a semi-supervised manner.

2.2. Learn to count with limited data

Relieving the crowd counting annotation efforts by using weakly/semi-/un-supervised learning mechanism has attracted researchers' attention for the past two years. For example, Liu *et al.* [19] leveraged a large number of unlabeled images and introduced a pairwise ranking loss to estimate the density map. Along the same line, Yang *et al.* [52] proposed a soft-label sorting network to regress the counting numbers rather than density map, which results in a performance reduction because of the difficult optimization from the input images to the target of scalar. Further, Wang *et al.*[49, 50] focused on a different direction, where they combined the synthetic images and realistic images to minimize the annotation burden. However, there is a domain gap between the synthetic and real-world scenarios; thus, they need further manual selections from the synthetic data. More recently, pseudo-labeling based semi-supervised approaches [40, 21] estimated the pseudo-ground truth of the unlabeled data, which is then used to supervise the network and improve the performance. Similarly, active learning-based methods [60, 20] annotated the most informative images instead of the whole training dataset and learned to count upon them. These methods can be effectively performed on the unlabeled data, but the model may be misled by the inevitable noisy supervision from the unlabeled data due to the aleatoric uncertainties [27], such as massive occlusions, complex backgrounds, *etc*.

2.3. Most Related Works

The framework of the most recent state-of-the-art method [21] is shown in Fig.1 (a), where the surrogate task (binary segmentation) learning mechanism is used to learn a robust feature extractor in a semi-supervised manner. We believe that learning a better feature extractor can be more reliable towards the unlabeled data's noisy supervision. However, there are some fundamental limitations in their framework: (1) The unlabeled data are only used to train the feature extractor and the binary segmentation predictor, aiming to avoid noise from unlabeled data contaminating the density regressor. However, it also leads to a side effect that only limited labeled data is used to train the density map predictor, subject to sub-optimal results. (2) Due to the unlabeled data's inevitable inherent noise, their model may provide incorrect predictions with spuriously high confidence because of the noisy supervision. This challenge has also been observed in other weakly/semi-/un-supervised crowd counting methods [30, 45, 40, 19, 52]. (3) The inherent prediction perturbation on spatial regions between the binary segmentation task and the density regression task may mislead the feature extractor's feature learning. In other words, the spatial inconsistency exists in the binary segmentation and density regression task.

We propose a semi-supervised model to address all the limitations mentioned above, and a simplified diagram of the model is shown in Fig.1 (c). Firstly, we introduced novel 'hard' uncertainty and 'soft' uncertainty from the teacher model to assist the student network to learn high-confident binary segmentation and density map predictions of the unlabeled data. This can alleviate the inevitable noisy supervision from the unlabeled dataset. Secondly, we proposed a novel differentiable transformation layer that converts the predicted density maps into approximated binary segmentation maps, where the inherent consistency loss is employed to avoid the prediction perturbations issues. Thirdly, because of the proposed uncertainty map and inherent consistency regularization, the feature extractor, binary segmentation predictor and density regressor in the student model of our work can benefit from both the labeled and unlabeled data and avoid sub-optimal issues; details of the proposed components are explained in the following sections.

3. Methods

The ground truth of the density map is generated by [12]. The binary segmentation ground truth mask is generated from the density map ground truth. Specifically, the value for each pixel in the binary ground truth mask is set to 1, if the pixel value of the density map is greater than 0, and 0, otherwise.

The proposed Teacher-Student framework structure is illustrated in Fig.2. The uncertainty map is estimated from the surrogate task with unlabeled data in the teacher model. Then we use the uncertainty map to assist the surrogate and the main task feature learning in the student models. The inherent consistency regularization between the surrogate task (binary segmentation) and the main task (density regression) in the student model improves its robustness regarding task-level spatial crowd region consistency.

3.1. Uncertainty Map Estimation

Different from the recent fully-supervised Gaussian distribution uncertainty-based [9] crowd counting method [28, 27], we propose a semi-supervised method to estimate the spatial uncertainty from the surrogate task (binary segmentation) in the teacher model with the unlabeled data, then use the uncertainty to assist the binary segmentation and density regression tasks feature learning in the student model so as to address the noisy supervision. This design is motivated by three considerations: (1) For crowd counting, the inevitable noise exists in many scenes, such as massive occlusions, complex backgrounds, *etc.*, which results in uncertain crowd regions [27]. So, the guidance of the proposed spatial uncertainty from the binary segmentation can be essential to alleviate the effects of noise. (2) Without the annotations in the unlabeled inputs, the predicted outputs from the teacher model may be unreliable and noisy. Therefore, an uncertainty-aware learning scheme is essential for the student model to assess the uncertainty and conduct a more reliable consistent feature learning. (3) The uncertainty estimated from the binary segmentation task indicates the uncertain locations of the crowd, which should be considered in the density regression task. Because the non-crowd regions should only maintain zero pixel values in the density map, the density regressor may produce larger pixel values due to the unlabeled data's spatial noise.

Recent domain adaptation studies [23, 46, 61] indicated that due to the domain gap, the models trained on source domain tend to produce under-confident, *i.e.* high-entropy predictions on the target domain. We found that such a phenomenon also exists in semi-supervised crowd counting tasks. Specifically, in our model, the outputs of the binary segmentation with unlabeled data in the teacher model tend to produce under-confident regions (the boundary along crowd regions). As mentioned in *Section.* 1 , this is because of the inevitable noise of the unlabeled data. Please refer to Fig. 3 for the qualitative uncertainty visualisation. To address this challenge, we adopt Shannon Entropy [32] as the metric to measure the randomness of the information [31], which is referred to as the uncertainty in this work. We then propose the 'hard' and 'soft' uncertainty maps to purify the learning process with the unlabeled data. Formally, given a C-dimensional softmax predicted class score $P_x^{(H,W,C)}$ from a $H \times W$ input image x, the Shannon Entropy $I_x^{(H,W)}$ is defined as:

$$I_x^{(H,W)} = -\sum_{c=1}^{C} P_x^{(H,W,C)} \odot \log P_x^{(H,W,C)}, \qquad (1)$$

where $\odot$ is Hadamard Product; C is the number of classes, which is 2 in our work because of the binary segmentation. In practice, we perform T times stochastic forward passes on the teacher model under random dropout and Gaussian noise input for each unlabeled input image. Therefore, we obtain a set of softmax probability vectors: $\{P^t\}_{t=1}^{T}$ from the segmentation branch, then the predicted class score $P^{(H,W,C)}$ is equal to $\frac{1}{T}\sum_{t=1}^{T} P^t$, thus we can obtain $I^{(H,W)}$ with equation 1.

With the assistance of the approximated Shannon Entropy $I^{(H,W)}$, we design two strategies to address the spatial uncertainty upon binary segmentation and density regression tasks between the student model and the teacher model, respectively. Firstly, the 'hard' uncertainty map U_h is introduced to guide the consistency learning on binary segmentation. In detail, we set a *threshold* and filter out the relatively unreliable binary segmentation predictions of the teacher model and select only the certain predictions as the target for the student model to learn from. In practice, the 'hard' uncertainty map U_h is equal to $1(I_{H,W} < threshold)$, where $1(\cdot)$ is a indicator function. Secondly, the 'soft' uncertainty map U_s is proposed to assist the consistency learning on the density regression task. The uncertain crowd regions do contain noisy density map predictions; however, only relying on the spatial uncertainty and filtering out the uncertain density map predictions may mislead the density regression or even add more noises. In addition to the spatial uncertainty, there are also other uncertainties caused by perspective distortions, non-uniform distribution, weather changes, *etc.*. Our ablation study experiments prove that the 'soft' uncertainty maps are more friendly than 'hard' uncertainty maps regarding to the consistency learning upon density map regression, thus advancing a performance boost. We retain all the density map predictions of the teacher model and introduce the 'soft' uncertainty map as a weighted mask to assign different weights to each pixel on the density map prediction according to the spatial certainty level. In detail, for these relatively reliable regions (pixels), we assign them more weights during the training to enforce the consistent learning to focus on

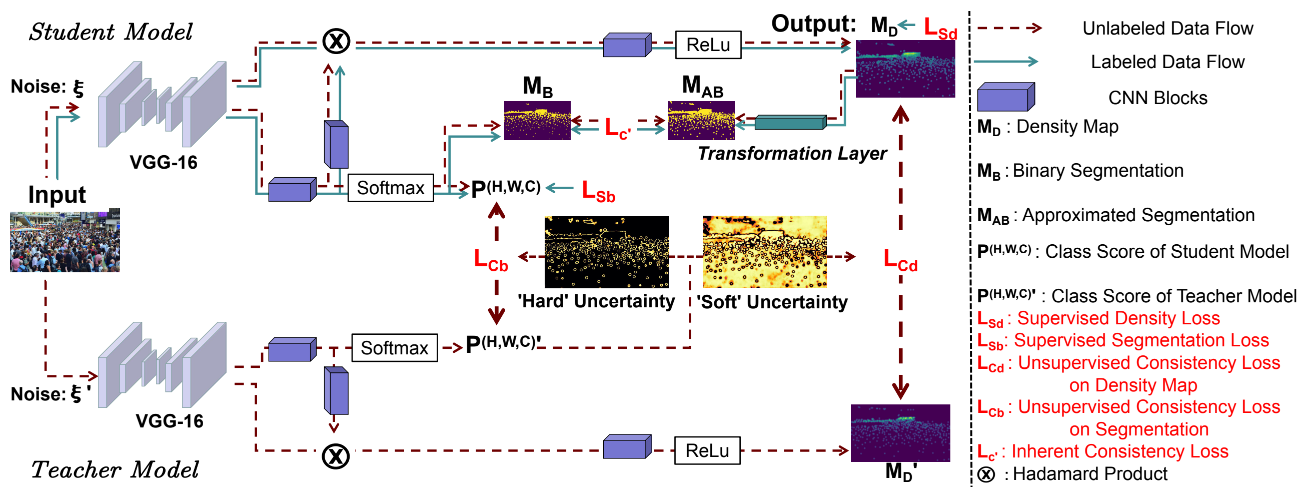

Figure 2. The pipeline of our uncertainty-aware framework for semi-supervised crowd counting enabled by the regularized surrogate task. The student model is optimized by minimizing the supervised density regression loss L_{Sd}, the binary segmentation loss L_{Sb} on the labeled data; the unsupervised inherent consistency loss $L_{c'}$ on both the unlabeled data and labeled data, the unsupervised consistency loss L_{Cb} and L_{Cd} on the unlabeled data. The estimated spatial uncertainty ('hard' and 'soft' uncertainty maps) from the teacher model guides the student to learn more reliable targets from the teacher. $P^{(H,W,C)\prime}$ and $M_D{}'$ are the outputs of the predicted class score and density map from the teacher model, which serves as the targets for the student model to learn from through consistency loss L_{Cb} and L_{Cd} respectively.

the certain prediction regions, while the relatively uncertain regions are still involved during the training with lower weights. In practice, we normalize the estimated Shannon Entropy $I^{(H,W)}$ into range $(0, 1)$ as $\hat{I}^{(H,W)}$, then define the 'soft' uncertainty map U_s as: $U_s = M * (1 - \hat{I}^{(H,W)})$, where M is the constant value of weighted mask to control the U_s pixel values.

With the estimated 'hard' and 'soft' uncertainty map, the uncertainty-aware unsupervised consistency loss (L_{Cd} & L_{Cb}) upon the main task (density regression) and surrogate task (binary segmentation) between the teacher and the student models can be guided during the training. Details will be shown in *Section.* 3.3.

3.2. Regularized Surrogate Task via Transformation Layer

In Fig.1 (a), recent method [21] utilized surrogate task to learn a robust feature extractor, which leads to the indirectly improved performance of density regressor. However, they did not consider the potential prediction perturbation on spatial regions due to the domain gap of feature learning from multi-tasks [25]; in which case, the binary segmentation can learn a different interest of the spatial regions compared with the one of the density regressor. To address the challenge, we proposed a simple yet effective differential transformation layer $\sigma(\cdot)$ to approximate the binary segmentation maps from the density regressor's output. In this way, we build a spatial regularization between the two tasks to address the potential inherent prediction perturbation issues in [21]. Meanwhile, the inherent consistency loss ($L_{c'}$) is employed between the binary segmentation predictions (M_B) and the approximated binary segmentation maps (M_{AB}) to regularize the surrogate task learning.Note that, $M_B \in \mathbb{R}^{H \times W \times 1}$ is the feature map of the corresponding crowd channel of the predicted class score $P^{(H,W,C)} \in \mathbb{R}^{H \times W \times 2}$.

Following the same process that we generate the binary segmentation ground truth mask from the density map ground truth, to convert the predicted density maps into approximated binary segmentation maps, an intuitive way is to use the Heaviside step function to set all the positive pixel values in the predicted density maps to 1 and zero pixel values to 0. However, it is impractical in training because of the non-differentiability. Hence, we proposed a simple yet effective differential transformation function to guarantee that purpose. With the output from the density regressor M_D, and the differential transformation layer $\sigma(\cdot)$, the approximated binary segmentation map M_{AB} is defined as:

$$M_{AB} = \sigma(K * M_D) = 2 * Sigmoid(K * M_D) - 1, \quad (2)$$

where K is a very large constant, which is set as 6000 in our work. Notably, as shown in Fig.2, M_D is a non-negative density map prediction because of the use of ReLu as the activation. In terms of such transformation function $\sigma(\cdot)$, the spatial consistency can be enforced between the two different tasks in a trainable manner. Specifically, the density regressor focuses on the pixel values regression, while the binary segmentation predictor aims for semantic and spatial reasoning. Thus, the natural task-level prediction difference on spatial crowd regions of these two tasks can be regularized by an unsupervised inherent consistency loss function $L_{c'}$ between the M_B and M_{AB}.

3.3. Loss Function

We optimize the student model using the supervised loss (density regression, binary segmentation) on the labeled data and the unsupervised consistency loss on the unlabeled data. The whole network is end-to-end trainable, and the total loss function L_{total} comprising five loss terms:

$$L_{total} = L_{Sd} + \alpha \cdot L_{Sb} + L_{c'} + \lambda \cdot (\alpha \cdot U_h \odot L_{Cb} + \cdot U_s \odot L_{Cd}), \quad (3)$$

where $\odot$ is Hadamard Product, L_2 loss is used for the supervised density map regression L_{Sd}; categorical cross-entropy loss is used for supervised binary segmentation L_{Sb} in the student model. Besides, α is a hyper-parameter to trade-off between the main task (density regression) and surrogate task (binary segmentation), which is set as 0.1 in our work. As for the unsupervised consistency loss, firstly, L_2 loss is used for unsupervised inherent consistency loss $L_{c'}$ between the binary segmentation predictions and the approximated binary segmentation maps from density map predictions in the student model; secondly, 'hard' uncertainty map U_h is used to assist the unsupervised consistency loss L_{Cb} upon the binary segmentation and 'soft' uncertainty map U_s is used for unsupervised density map regression consistency loss L_{Cd}. Here, we choose Euclidean distance as the consistency metric for L_{Cd} and L_{Cb}. λ are adopted from [11] as the same time-dependent Gaussian ramp-up weighting coefficient to trade-off between the supervised loss and unsupervised loss. This is to avoid the network get stuck in a degenerate solution, where no meaningful prediction of the unlabeled data is obtained [11].

4. Experiments

4.1. Data sets

ShanghaiTech [58] consists of 1198 images, containing a total amount of 330165 people with head centre point annotations. This dataset has two parts: **SHA** includes 482 images and is divided into a training (300) and testing (182) subset. **SHB** includes 716 images and is divided into 400 images for training and 316 images for testing. **UCF-QNRF** [6] is a large crowd dataset, consisting of 1,535 images with about 1.25 million annotations in total. As indicated by [6], 1201 images are used for training; the remaining 334 images form the test set. **JHU-Crowd** [41] is a recent challenging large-scale dataset that containing 4372 images with 1.51 million annotations. This dataset is divided into 2272 images for training, 500 images for validation, and 1600 images for testing. **NWPU-Crowd** [48] is current the largest public crowd counting dataset, containing 5109 images with over 2.13 million annotations. The dataset includes 3109 training images and 500 validation images; due to no access to the testing images; instead, we keep their validation images to evaluate our model's performance. Note that, we set 50% of the training images as the labeled data and the rest as the unlabeled data. In particular, for ShanghaiTech (part A, part B), UCF-QNRF and NWPU-Crowd, we use 10% of the labeled training images as the validation dataset.

4.2. Implementation Details

Code is available at: `https://anonymous.4open.science/r/d8fcc5eb-57a6-4b78-a332-db3024499e6b/` We adopt a truncated VGG-16 [37] as the backbone network, which is the same as [18, 13, 21, 44, 19]. Additionally, following [44], two dropout layers with a drop out rate of 0.5 are added into the feature extractor to introduce model-level perturbations. The dropout is turned on during the training and turned off during the testing. Please refer to supplementary for detailed model structure. We update the teacher model's weight θ' as an EMA of the student model's weight θ during the training step, such as $\theta'_t = \zeta \cdot \theta'_{t-1} + (1-\zeta) \cdot \theta_t$, where t is the t_{th} training step, and ζ is the EMA decay to control the update rate, which is empirically set as 0.999 in our work. For Shannon Entropy estimation, we set $T = 8$ as the stochastic forward passes times to balance the model's performance and training efficiency. Besides, we set the *threshold* as a Gaussian ramp-up function from 3/4 maximum uncertainty value to maximum uncertainty value for' hard' uncertainty map estimation. For the 'soft' uncertainty map estimation, the weight value M is set as 7. Details of the hyper-parameter setting in our work can be found in the supplement.

The training data set is augmented by randomly cropping the input images, the density maps ground truth, and the binary segmentation ground truth with fixed size 128×128 at a random location; then randomly horizontal flipped the image patches with the probability of 0.3. We trained our model up to 600 epochs or stop early when the network has converged, with an initial learning rate of 7e-5 and divided by 5 every 200 epochs. The batch size is set as 16, consisting of 8 labeled images and 8 unlabeled images. All the training processes are performed on a server with 8 TESLA V100, and all the testing experiments are conducted on a local workstation with a Geforce RTX 2080Ti.

5. Results

In this section, we present our experimental results on the crowd counting tasks compared to previous state-of-the-art methods. Following the previous methods, we adopt Mean Absolute Error (MAE) and Root Mean Squared Error $(RMSE)$ to evaluate the counting performance. The results of ablation study are also shown to demonstrate the importance of the various components in our framework, such as the number of labeled and unlabeled images, 'soft' and

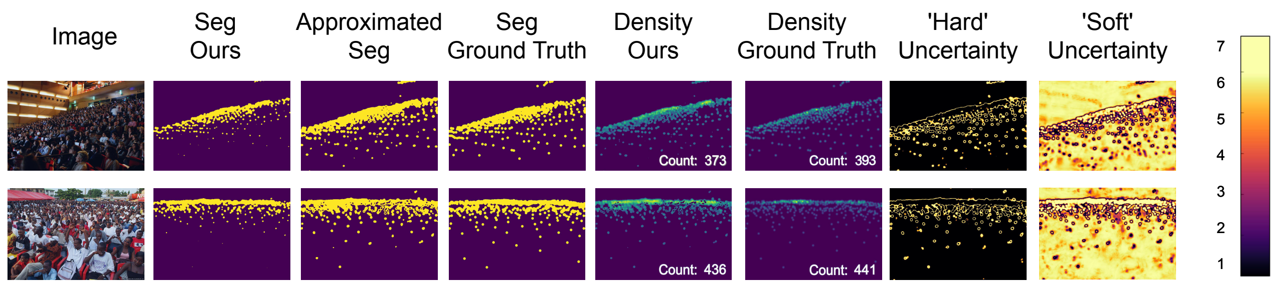

Figure 3. Qualitative results on SHA test dataset. In the ‘hard’ uncertainty maps, the yellow pixels represent uncertain regions and the black pixels are certain regions. In the ‘soft’ uncertainty maps, the different color represents different weight mask values according to the color bar; higher value denotes more certain regions. The estimated ‘soft’ uncertainty indicates that the crowd regions’ boundary is more uncertain than other regions, which is reasonable because of the complex backgrounds.

Methods		SHA		SHB		QNRF		JHU-Crowd		NWPU-Crowd	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Fully-Supervised	CACC [18]	62.3	100.0	7.8	12.2	107.0	183.0	100.1	314.0	93.6	489.9
	CSR-Net [13]	68.2	115.0	10.6	16.0	119.2	211.4	85.9	309.2	104.9	433.5
	Ours (Fully)	66.9	125.6	12.3	17.9	119.2	213.3	80.1	305.3	105.8	445.3
Semi-supervised	Mean-Teacher [44] (Baseline)	88.2	151.1	15.9	25.7	147.2	249.6	121.5	388.9	129.8	515.0
	L2R [19]	86.5	148.2	16.8	25.1	145.1	256.1	123.6	376.1	125.0	501.9
	Sindagi *et al.* [40]	89.0	-	-	-	136.0	-	-	-	-	-
	Liu *et al.* [21]	-	-	-	-	138.9	-	-	-	-	-
Ours (Label-Only)		94.6	152.0	19.2	31.9	152.9	266.1	133.3	415.0	141.0	625.6
Ours (Semi)		**68.5**	**121.9**	**14.1**	**20.6**	**130.3**	**226.3**	**80.7**	**290.8**	**111.7**	**443.2**

Table 1. Quantitative results on four crowd counting datasets. Our model achieves superior performance than the other semi-supervised methods in terms of MAE with the same setting of 50% labeled data on four datasets.

‘hard’ uncertainty maps, differential transformation layer, respectively. Quantitative results are shown in Tab.1, 2, 3 and Fig.4. Fig. 3 show the qualitative results. More qualitative results can be found in the supplementary. More quantitative results compared with previous methods ([21, 40, 44]) under different number of labeled data settings are shown in the supplementary.

5.1. Crowd Counting Results

Fig.3 shows qualitative results; specifically, we present the predicted and approximated segmentation maps, and the visualized uncertainty maps to demonstrate our model’s cohesion, along with the contribution of spatial uncertainty guidance and inherent consistency regularization. In particular, we compare our model with previous semi-supervised methods [19, 21, 40, 44]. The results of [21, 40] are retrieved from their published papers, and we re-implement the rest methods [19, 44] through running their public code. Note that, [21] adopts the same backbone (VGG-16 [37]) as our model; they build their model based on CSR-Net [13], which achieves a comparable performance under fully-supervised manner with ours (*i.e. Ours (Fully) in the Tab. 1*). [40] adopts a more powerful backbone producing superior performance than *Ours (Fully)* under fully-supervised manner. So the comparison with them in a semi-supervised manner can be seen as straightforward and reasonable. Additionally, we add binary segmentation module into the Baseline model [44] to maintain similar model parameters as Ours (Semi); however, without the proposed transformation layer and uncertainty maps, the Baseline model achieves relatively 18.5 % higher MAE compared with Ours (Semi) on four datasets. To make an intuitive comparison, we also present different prediction results with our proposed model: (1) Ours (Label-Only): trained with half labeled data on the student model (without transformation layer). (2) Ours (Semi): trained with half labeled and half unlabeled data on the student and teacher model simultaneously; inferred with student model only. (3) Ours (Fully): trained with all the labeled data on the student model (without transformation layer). Note that, the transformation layer works as an activation function, which hardly increases the size of the model. Tab.1 shows that Ours (Semi) outperforms the Ours (Label-Only) by a large margin with average 25.1% performance gain in terms of MAE on four datasets, which is benefits from the proposed uncertainty maps, differential transformation layer and unlabeled data. In particular, our model achieves comparable performance with only 50% labeled data, compared with Ours (Fully) with 100% labeled data in SHA and JHU-Crowd dataset. Furthermore, to present comprehensive comparisons, we also show the performance of previous state-of-the-art crowd counting methods [13, 18] with the same backbone network as ours under a fully supervised manner. Tab.1 shows our method outperforms other semi-supervised methods in terms of MAE and RMSE on all four datasets under the same test settings and achieves a comparable performance to the previous state-of-the-art fully supervised works in SHA and JHU-Crowd dataset.

5.2. Ablation Study

We investigate the effect of each component in our proposed model. Our model is robust to the hyper-parameters; results of more ablation studies, such as coefficients of the loss function, *threshold* of 'hard' uncertainty map, weights of 'soft' uncertainty map, *etc.*, can be found in the supplementary.

Ablation on Number of Labeled & Unlabeled Images: We examine the performance of Baseline [44] and Ours (Semi) with a different number of labeled & unlabeled images. We conduct experiments on the SHA dataset by varying the number of labeled images from 30 to 150 while fixing the number of unlabeled images to be 150; or varying the number of unlabeled images from 30 to 150 while fixing the amount of labeled images to be 150. The performance are shown in Fig.4, where it shows Ours (Semi) achieves consistent superior performance over the Baseline [44], which demonstrate the robustness of our method.

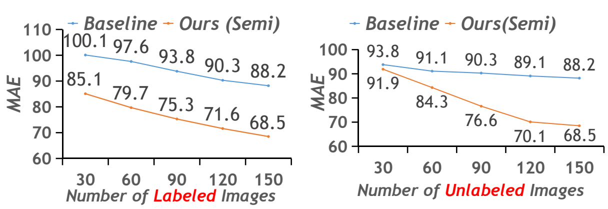

Figure 4. The impact of the number of labeled & unlabeled images. Evaluated on SHA dataset in terms of MAE.

Ablation on Uncertainty Map: We conduct several experiments to evaluate the impact of the proposed uncertainty maps (Unc). Firstly, we remove both the 'hard' and 'soft' uncertainty maps and keep the rest model structure. Notably, the concept of 'surrogate task' is used for spatial uncertainty estimation from binary segmentation task in this work; if we remove the uncertainty module, the binary segmentation task will only be served as information supplement for intermediate feature learning. Secondly, we add either 'hard' uncertainty map or 'soft' uncertainty map respectively to evaluate the effectiveness of each of them. Thirdly, we add two 'hard' uncertainty maps to verify the effectiveness of the proposed 'soft' uncertainty map with respect to the consistency learning on the density regression. Finally, we add both 'hard' and 'soft' uncertainty maps (Ours) for further comparison. Tab.2 shows that our model with both uncertainty maps achieves average 15.5% and 16.0% performance gain via MAE compared with that without uncertainty map employed on SHA and JHU-Crowd datasets, respectively. This proves that our proposed uncertainty maps can assist the feature learning between the student and teacher model and further improve the performance.

Methods	SHA		JHU-Crowd	
	MAE	RMSE	MAE	RMSE
w/o Unc	81.1	143.1	96.1	311.9
w/ 'Hard' Unc	77.3	137.0	92.7	304.0
w/ 'Soft' Unc	73.1	130.8	85.3	296.2
w/ two 'Hard' Unc	72.1	128.9	83.2	294.8
w/ both Unc (ours)	**68.5**	**121.9**	**80.7**	**290.8**

Table 2. Performance comparison of the effectiveness of the proposed uncertainty maps. Compared with the 'hard' uncertainty maps, the 'soft' uncertainty maps can bring average 6.5% superior performance improvement via MAE on two datasets.

Ablation on Transformation Layer: We perform several experiments to analyse the impact of the proposed transformation layer (Trans). In detail, we remove the transformation layer and inherent consistency loss ($L_{c'}$), and keep the rest components in our model. Then we employ the transformation layer, and $L_{c'}$ upon (1) labeled data only, (2) unlabeled data only, (3) both of the labeled and unlabeled data to demonstrate the performance gain. Tab.3 shows that applying transformation layer only on the unlabeled data can gain close results to ours, and only applying transformation layer on the labeled data results in average 4.4% performance decline via MAE on two datasets compared with ours. The above proves that the performance gain of our model in terms of the proposed differential transformation layer is mainly from the unlabeled data.

Methods	SHA		JHU-Crowd	
	MAE	RMSE	MAE	RMSE
w/o Trans	74.8	131.0	89.3	301.2
w/ Trans on Label	73.2	129.5	86.8	296.3
w/ Trans on unlabeled	70.7	123.9	82.1	293.4
w/ Trans on both (ours)	**68.5**	**121.9**	**80.7**	**290.8**

Table 3. Ablation study on the impact of the proposed differential transformation layer. When applying the transformation layer on both the unlabeled and labeled data, ours achieves average 9.1% performance gain than the model without transformation layer via MAE on two datasets.

6. Conclusions

We propose a spatial uncertainty-aware semi-supervised crowd counting methodology via regularized surrogate task to alleviate the inevitable noisy supervision from the unlabeled data. We have demonstrated its potentials in reducing annotations efforts while maintaining good performance upon four challenging crowd counting datasets. It is anticipated that our approach will be widely applicable in the real world.

References

[1] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4594–4603, 2020. 1

[2] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–127. Springer, 2019. 1

[3] Lokesh Boominathan, Srinivas SS Kruthiventi, and R Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 640–644, 2016. 3

[4] Junyu Gao, Qi Wang, and Xuelong Li. Pcc net: Perspective crowd counting via spatial convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 2, 3

[5] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2888–2897, 2019. 1

[6] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–546, 2018. 6

[7] 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. 1

[8] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4706–4715, 2020. 3

[9] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 1, 4

[10] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018. 1

[11] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 2, 6

[12] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in neural information processing systems*, pages 1324–1332, 2010. 4

[13] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018. 6, 7

[14] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1830, 2019. 3

[15] Chenchen Liu, Xinyu Weng, and Yadong Mu. Recurrent attentive zooming for joint crowd counting and precise localization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1217–1226. IEEE, 2019. 3

[16] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2018. 3

[17] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3225–3234, 2019. 1

[18] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, 2019. 1, 6, 7

[19] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7661–7669, 2018. 3, 6, 7

[20] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Exploiting unlabeled data in cnns by self-supervised learning to rank. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1862–1878, 2019. 1, 3

[21] Yan Liu, Lingqiao Liu, Peng Wang, Pingping Zhang, and Yinjie Lei. Semi-supervised crowd counting via self-training on surrogate tasks. *European Conference on Computer Vision*, 2020. 1, 2, 3, 5, 6, 7

[22] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. Point in, box out: Beyond counting persons in crowds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2019. 1

[23] Mingsheng Long, Yue Cao, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Transferable representation learning with deep adaptation networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):3071–3085, 2018. 4

[24] Ao Luo, Fan Yang, Xin Li, Dong Nie, Zhicheng Jiao, Shangchen Zhou, and Hong Cheng. Hybrid graph neural networks for crowd counting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 3

[25] Xiangde Luo, Jieneng Chen, Tao Song, Yinan Chen, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation through dual-task consistency. *AAAI Conference on Artificial Intelligence*, 2021. 2, 5

[26] Yunqi Miao, Zijia Lin, Guiguang Ding, and Jungong Han. Shallow feature based dense attention network for crowd counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11765–11772, 2020. 3

[27] Min-hwan Oh, Peder A Olsen, and Karthikeyan Natesan Ramamurthy. Crowd counting with decomposed uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11799–11806, 2020. 1, 2, 3, 4

[28] Viresh Ranjan, Boyu Wang, Mubarak Shah, and Minh Hoai. Uncertainty estimation and sample selection for crowd counting. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 4

[29] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and Venkatesh Babu Radhakrishnan. Locate, size and count: Accurately resolving people in dense crowds via detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[30] Deepak Babu Sam, Neeraj N Sajjan, Himanshu Maurya, and R Venkatesh Babu. Almost unsupervised learning for dense crowd counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8868–8875, 2019. 3

[31] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001. 4

[32] Claude Elwood Shannon and Warren Weaver. The mathematical theory of communication. univ. *Illinois press, Urbana, I*, 11:117, 1949. 4

[33] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5245–5254, 2018. 1

[34] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting perspective information for efficient crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7279–7288, 2019. 1

[35] Zenglin Shi, Pascal Mettes, and Cees GM Snoek. Counting with focus for free. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4200–4209, 2019. 2, 3

[36] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5382–5390, 2018. 1

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 6, 7

[38] Vishwanath A Sindagi and Vishal M Patel. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Transactions on Image Processing*, 29:323–335, 2019. 3

[39] Vishwanath A Sindagi and Vishal M Patel. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1002–1012, 2019. 1

[40] Vishwanath A Sindagi, Rajeev Yasarla, Deepak Sam Babu, R Venkatesh Babu, and Vishal M Patel. Learning to count in the crowd from limited labeled data. *European Conference on Computer Vision*, 2020. 1, 3, 7

[41] 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. 6

[42] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *arXiv preprint arXiv:2004.03597*, 2020. 1

[43] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 2

[44] 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*, pages 1195–1204, 2017. 2, 6, 7, 8

[45] 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*, pages 365–380. Springer, 2016. 3

[46] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2517–2526, 2019. 4

[47] Jia Wan, Wenhan Luo, Baoyuan Wu, Antoni B Chan, and Wei Liu. Residual regression with semantic prior for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4036–4045, 2019. 1

[48] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 6

[49] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8198–8207, 2019. 1, 3

[50] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Pixel-wise crowd understanding via synthetic data. *International Journal of Computer Vision*, pages 1–21, 2020. 1, 3

[51] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Reverse perspective network for perspective-aware object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4374–4383, 2020. 1

[52] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Weakly-supervised crowd counting learns from sorting rather than locations. In *European Conference on Computer Vision*, 2020. 3

[53] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019. 1

[54] Anran Zhang, Lei Yue, Jiayi Shen, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Attentional neural fields for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5714–5723, 2019. 3

[55] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 833–841, 2015. 3

[56] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020. 1

[57] Qi Zhang and Antoni B Chan. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8297–8306, 2019. 1

[58] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 3, 6

[59] Muming Zhao, Jian Zhang, Chongyang Zhang, and Wenjun Zhang. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12736–12745, 2019. 2, 3

[60] Zhen Zhao, Miaojing Shi, Xiaoxiao Zhao, and Li Li. Active crowd counting with limited supervision. *European Conference on Computer Vision*, 2020. 1, 3

[61] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019. 4