Calibrating Uncertainty for Semi-Supervised Crowd Counting

Chen Li * Xiaoling Hu Shahira Abousamra Chao Chen
Stony Brook University

Abstract

Semi-supervised crowd counting is an important yet challenging task. A popular approach is to iteratively generate pseudo-labels for unlabeled data and add them to the training set. The key is to use uncertainty to select reliable pseudo-labels. In this paper, we propose a novel method to calibrate model uncertainty for crowd counting. Our method takes a supervised uncertainty estimation strategy to train the model through a surrogate function. This ensures the uncertainty is well controlled throughout the training. We propose a matching-based patch-wise surrogate function to better approximate uncertainty for crowd counting tasks. The proposed method pays a sufficient amount of attention to details, while maintaining a proper granularity. Altogether our method is able to generate reliable uncertainty estimation, high quality pseudo-labels, and achieve state-of-the-art performance in semi-supervised crowd counting.

1. Introduction

Crowd counting is the task of estimating the number of individuals in images or videos. It is important for public security, transport management, crowd surveillance, and catastrophe management, to name a few. Deep-learning-based methods [51, 60, 63, 43, 41, 1, 37, 11, 24, 40, 39, 25] have achieved promising performance in crowd counting tasks. However, to achieve superior performance, these methods require a large amount of annotations. For each image, hundreds or thousands of points/markers are added to indicate pedestrians (see Fig. 1(b)). Acquiring such annotations is costly and time-consuming; it takes 2000 human hours to annotate UCF-QNRF [10] dataset, which contains 1.25 million points/markers.

To alleviate the annotation burden, semi-supervised crowd counting is explored to leverage the information in unlabeled images with fewer annotated images. Previous works use self-supervised learning [26] and data synthesis [54, 55]. Another promising direction is *pseudo-labeling*, i.e., generating pseudo-labels for unlabeled data based on model prediction, and adding them as additional training data. Pseudo-labeling has been used in many other vision tasks, such as classification [47], segmentation [4], and object detection [58]. To select reliable predictions as pseudo-labels, one may use the *uncertainty* of the model. Predictions with low uncertainties are considered more reliable and can be used as pseudo-labels to train the model.

Despite the wide application of uncertainty in computer vision [36, 5, 17, 15, 3, 62, 58, 8, 6], estimating uncertainty in crowd counting remains a challenge, due to the heterogeneity of the marker distributions and fundamental issues of a crowded scene including perspective, occlusion, and cluttering. A few previous works [34, 29] estimate uncertainty for crowded scenes through the consistency between predictions made by different models. However, these methods completely rely on models' predictions on unlabeled data. Without a proper controlling mechanism, the uncertainty estimation can not be guaranteed to be reliable. In challenging regions, where all models make similar mistakes, consistency may create overconfidence yet wrong predictions. The training may be derailed by noisy pseudo-labels.

In this paper, we propose a novel semi-supervised crowd counting approach. We focus on uncertainty estimation and propose *the first supervised solution for crowd counting*. Unlike previous methods, we use the labeled data as direct supervision for uncertainty estimation. Although the true uncertainty is unknown even for labeled data, we propose a surrogate function based on the similarity between model predictions and ground truth. Intuitively, when a model can predict well, it should be less uncertain about the data. This learning-performance-based surrogate function provides the opportunity to directly train the model for uncertainty estimation. One appealing feature is that the supervised uncertainty estimation is well under-control throughout different training stages, providing necessary stability to the dynamic semi-supervised learning process.

One critical question is the design of the surrogate function, i.e., how to measure the similarity between a prediction

*Email: Chen Li (li.chen.8@stonybrook.edu).

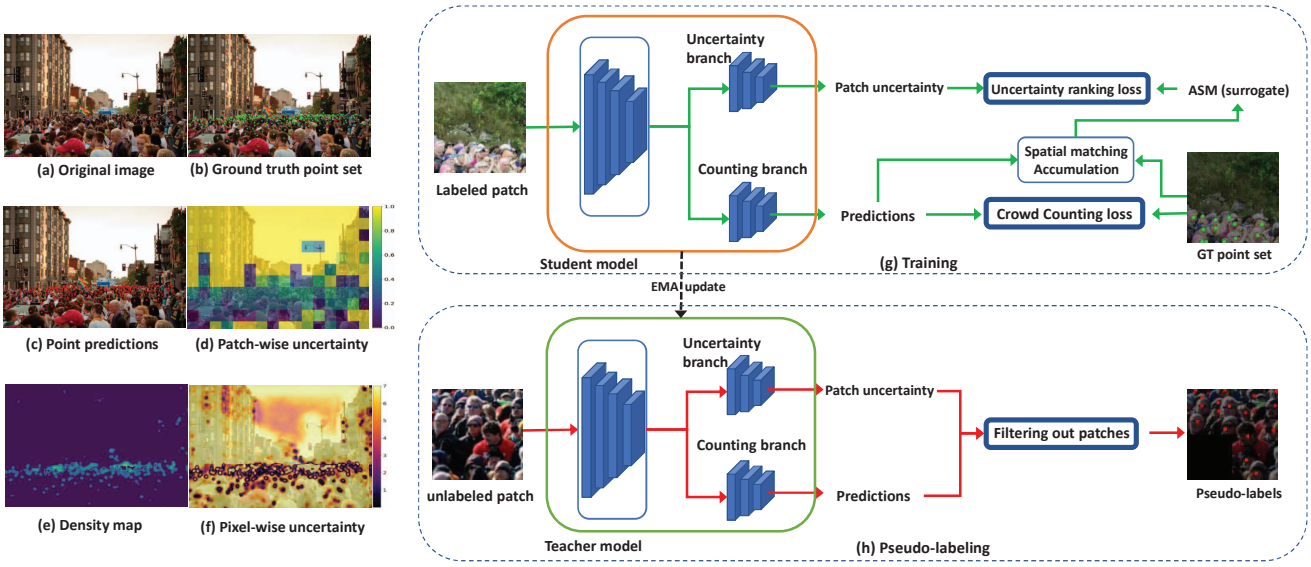


Figure 1: An overview of our method. The light color represents low uncertainty in the whole context. **(a)** Original image. **(b)** Ground truth point set. **(c)** Point predictions from our method. **(d)** Our patch-wise uncertainty map. **(e)** Density map generated by density-based method [34]. **(f)** Pixel-wise uncertainty generated by comparing density maps [34]. **(g)** The pipeline of student model training for crowd counting and for uncertainty estimation. A labeled patch goes through the feature extractor and uncertainty branch to generate an uncertainty score. The uncertainty score is compared with the surrogate function (ASM) using uncertainty ranking loss. Counting predictions generated by the counting branch are supervised by crowd counting loss. **(h)** Pseudo label generation process. We use a teacher model which is acquired from the student model via the EMA strategy. It generates predictions and patch uncertainty for all unlabeled patches. Patches with low uncertainty patches are used to supervise the training of crowd counting.

and the ground truth. As the crowd counting prediction can be evaluated in many different ways, a few different similarity measures have been used for uncertainty estimation. Direct evaluation of uncertainty at image level [65] cannot account for the heterogeneous spatial distribution of pedestrians. At the other extreme, Meng et al. [34] propose pixel-level uncertainty using the density maps, i.e., kernel density based on the points/markers (Fig. 1(e)). They compare the density maps of teacher and student models pixel-by-pixel to estimate uncertainty at every pixel. This uncertainty, however, can be unreliable because the density maps lack the necessary details, i.e., exact locations of points/markers. As shown in Fig. 1(f), one obtains very low pixel-wise uncertainty at high density regions, where the models tend to be erroneous; but the sky regions get higher uncertainty.

To tackle this challenge, our second contribution is a novel measure of uncertainty that better suits crowd counting tasks. We focus on the exact coordinates of points/markers and compare the prediction point set and ground truth point set (the set constructed by the coordinates of individuals) through a matching-based similarity metric, called Accumulated Spatial Matching distance (ASM). This matching-based strategy accounts for the full details of the prediction and avoids systematic over-count/under-count bias. To provide the proper granularity and account for the spatial heterogeneity of uncertainty, we partition images into small patches and estimate uncertainty at the patch level. Fig. 1(d) illustrates the patch-level uncertainty estimation by our method; the estimated uncertainty is high in crowded regions and low in regions with fewer markers, e.g., sky. This is consistent with our expectations.

We incorporate our supervised uncertainty estimation into an end-to-end semi-supervised learning pipeline. Our model uses a student model and a teacher model. During each training epoch, the student model is trained on the training set for both prediction and uncertainty estimation. The uncertainty training is by comparing with the proposed ASM surrogate function. The teacher model is a stabilized version of the student model; its weights are an exponential moving average (EMA) of the weights of the student model. The teacher model makes predictions and estimates uncertainties on unlabeled patches. Predictions with low uncertainty are chosen to be pseudo-labels and be used to supervise crowd counting.

In summary, our contribution is three-fold.

- We propose a novel supervised approach to calibrate patch-wise uncertainty for crowd counting task. The estimated uncertainty can be used to effectively se-

lect reliable pseudo-labels to enhance semi-supervised training.

- We propose a novel patch-wise matching-based similarity measure as a surrogate function for uncertainty. This surrogate function focuses on specific coordinates and cardinalities of points/markers and provides reliable information throughout different training stages.
- On various benchmarks, experimental results show that our method generates well-calibrated uncertainties, high-quality pseudo-labels, and achieves state-of-the-art performance on the semi-supervised crowd counting task.

2. Related Works

Deep learning based algorithms have achieved great progress in crowd counting tasks. In this section, we will review the works in crowd counting, uncertainty estimation, and semi-supervised learning.

Crowd counting. Density-based crowd counting frameworks are widely studied. The crowd counting problem is treated as a density estimation problem by density-based methods. The crowd size is calculated by summing all the pixel values of the density map. The density map can be generated pixel-by-pixel [18, 9, 28, 2, 12, 35] or patch-by-patch [57, 22, 23]. Wang et al. [52] utilize Optimal Transport (OT) to measure the distribution difference between the predicted density map and ground truth. Ma et al. [32] construct the likelihood function of individuals based on Gaussian distribution. Bai et al. [2] address the noise in crowd counting annotations and conduct a self-correction (SC) supervision framework, which can correct annotations based on the model's output. Many promising methods have been published in this track. Despite the strong performance of density-based methods, they cannot generate the accurate locations of individuals out of the predicted density map. Localization-based methods count the crowd by locating all individuals. Some of them utilize object detection techniques to get the individual locations [38, 30, 19]. However, most crowd counting datasets are only point annotated. This makes it difficult to acquire the precise coordinates of bounding boxes and leads to inferior model performance. The locations of individuals are also captured by dots [21] or blobs [16], but ad-hoc post-process is used to eliminate false-positive and separate joint individuals. Song et al. [48] propose a one-to-one matching framework to match the point proposals with ground truth locations. Instead of comparing the location distribution difference, they focus on finding positive proposals and increasing model confidence in these proposals.

Counting with limited annotations. It is laborious and time-consuming to annotate images for crowding counting tasks. Semi-/weakly supervised learning methods have been proposed to alleviate the annotation burden. Liu et al. [26] introduce a learning-to-rank framework to leverage the unlabeled information contained in the relation between crop size and counting number (image patches should include more individuals than their sub-patches). Yang et al. [61] propose a soft-label sorting network to utilize the counting information of crowd numbers rather than the locations of individuals. Xu et al. [59] propose a density-based framework for partial annotation setting. Sindagi et al [44] use an iterative learning framework with Gaussian Process (GP) to leverage unlabeled information and boost model performance. Liu et al. [29] construct a series of surrogate tasks for training a feature extractor on both labeled and unlabeled data with a self-training framework. Liu et al. [20] use a contrastive loss to intensify the learning of density maps on labeled and unlabeled data through density agency. Most previous methods can utilize unlabeled data properly, but the inherent noise in unlabeled supervision can deteriorate the model performance significantly.

Uncertainty estimation for crowd counting. Uncertainty estimation is extensively studied for tasks like segmentation [13, 33, 15, 3, 17] and object detection [7, 31]. For crowd counting, uncertainty estimation is important but understudied. Meng et al. [34] estimate pixel-wise uncertainty by comparing density maps generated by teacher and student models. As illustrated in Fig. 1(f), the pixel-wise uncertainty can be incorrect, especially in dense areas. Comparing models' output also makes the uncertainty less reliable. This method also treats the problem as a segmentation task and compares segmentation maps between teacher and student models for uncertainty estimation. This additional uncertainty map suffers from the same issues as the density-based uncertainty map. Zhao et al. [65] introduce uncertainty by comparing the crowd density distribution between different images and claim that the images with crowd dense distribution more dissimilar from labeled images have higher uncertainty. Liu et al. [27] design a self-supervised proxy task based on the fact that for the same region, the wider range scene should contain more individuals. The number of mistakes made on this proxy task can be seen as an uncertainty estimation. Our method differs from these previous methods in that it is the first one to use a supervised strategy for uncertainty calibration, ensuring a much more stable uncertainty map during training. Furthermore, our method focuses on the locations and cardinalities of markers and provides patch-level uncertainty maps. This ensures that important details are used while proper granularity is used.

3. Method

We propose a pseudo-labeling-based semi-supervised crowd counting approach using patch-level uncertainty estimation. Given a small set of labeled images and a large set

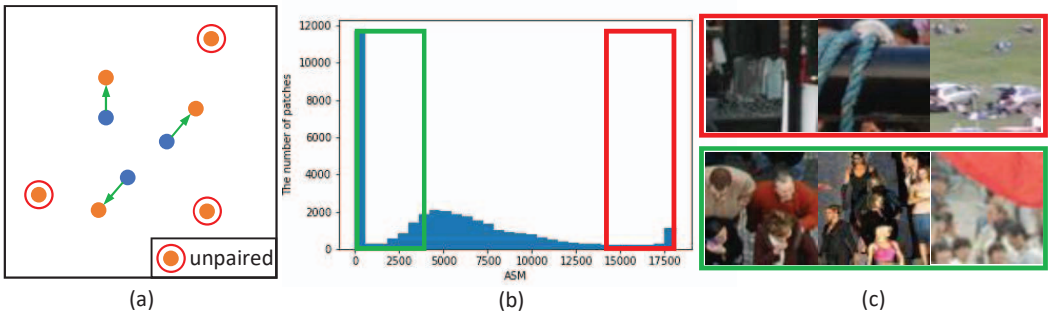


Figure 2: **(a)** An illustration of the introduction of the spatial matching distance. We pair all blue points with orange points one-by-one and use sums of the lengths of the green arrows as part of the matching distance. Three unpaired orange points will contribute a constant penalty each. **(b)** The ASM distribution on the ShanghaiTech Part-A training set. **(c)** Sample image patches from ShanghaiTech Part-A, have high ASM (red loop, top) and low ASM (green loop, bottom). The individuals in the red loop are difficult to locate while the crowds in the green loop are easy to count.

of unlabeled images, semi-supervised crowd counting is to train models with both labeled and unlabeled images. The key is to make the best use of unlabeled images. To this end, one effective approach is pseudo-labeling. A pseudo-labeling method assigns to unlabeled data "pseudo-labels", which are often model predictions. These pseudo-labeled data will be added to the training set to improve the model. The improved model will then produce better and more pseudo-labels. The model training and pseudo-label generation can enhance each other in a mutual manner.

The challenge of the above process is to decide which pseudo-labels are trustworthy. As the model is imperfect, its predictions can be noisy. Picking up too many pseudo-labels will inject noise and derail the training. On the other hand, an overly conservative strategy may not fully utilize the unlabeled data and cannot maximize the potential of the model. A useful measure for selecting pseudo-labels is uncertainty [47]. One may choose model predictions with lower uncertainty, hoping that these "certain" predictions are more likely to be correct.

In this paper, we propose a novel uncertainty estimation method for the crowd counting problem and incorporate it into the semi-supervised learning framework. In Sec. 3.1, we propose to explicitly train uncertainty estimation using labeled data. This way, the uncertainty is more reliable and can be used to select high-quality pseudo-labels. The training is based on a surrogate function that approximates the true uncertainty. We propose a novel surrogate function for crowd counting in Sec. 3.2. Finally, we introduce the semi-supervised learning framework in Sec. 3.3.

3.1. Uncertainty Calibration Using Labeled Data

Our method estimates uncertainty through supervision. This is in contrast with previous semi-supervised crowd counting works [34, 29], which directly calculate uncertainty on unlabeled data based on model prediction consistency. These previous methods, without direct supervision, can make unreliable uncertainty estimations without being noticed, leading to sub-optimal models.

An overview of our method is illustrated in Fig. 1(g). Our model takes an image patch as input. It has two branches: one for prediction, and one for uncertainty estimation. The two branches share the same feature extractor; this ensures that the uncertainty estimation is based on reliable high-level information and thus generalizes well to unseen data. During training, our model not only predicts on each unlabeled patch, but also generates uncertainty values for patches. The predictions with low uncertainty will be chosen as pseudo-labels and be added to the training set to refine the model. To ensure the quality of the generated uncertainty, we use labeled data as supervision to train the uncertainty estimation branch.

The key question is to find an appropriate measure of uncertainty on labeled data. In other words, we need a good "surrogate function" as the "ground truth" uncertainty. We draw inspiration from previous work on the classification task. Moon et al. [36] introduce an uncertainty surrogate function by counting the frequency of correct predictions of snapshot models at different training epochs. The intuition is that a model should be more uncertain on a sample if it is often misclassified. When it comes to the crowd counting task, we need to design a special metric to measure how "correct" a prediction is. To this end, we propose a novel similarity measure between ground truth and prediction. We accumulate the similarity across different training epochs. Details of the proposed surrogate function, called Accumulated Spatial Matching Distance (ASM), will be presented in Sec. 3.2.

As illustrated in Fig. 1(g), during model training, we not only provide supervision on model prediction, but also supervise the uncertainty branch by comparing its output with the surrogate function. This will ensure that the model can

be generalized well to unlabeled data, and produces reliable uncertainty for pseudo-label selection.

Pairwise ranking loss for uncertainty supervision. To train the uncertainty branch, we propose a pairwise ranking loss. The loss ensures that the rank of uncertainty output is consistent with the rank of uncertainty surrogate. Compared with the numeric values of the surrogate function, the ranking is more stable and reliable. This is why ranking loss can achieve better performance than pointwise loss, e.g., L1 loss. In particular, we use lambdaloss [56] as follows:

$$\mathcal{L}_{uncer} = \sum_{a_i > a_j} |a_i - a_j| \log_2(1 + e^{-(\kappa_i - \kappa_j)}) \quad (1)$$

where a_i is the batch min-max normalized ASM on patch I_i, and κ_i is the model's uncertainty output on patch I_i. The loss will incur a large penalty for a pair of patches with $a_i > a_j$ yet $\kappa_i < \kappa_j$. The loss is also weighted by $|a_i - a_j|$, i.e., the absolute ASM difference between i and j.

3.2. Surrogate for Uncertainty Calibration

In this section, we propose a novel measure of similarity between prediction and ground truth for each training patch. Since our focus is to measure the quality of pedestrian point set predictions, we cannot directly use previous similarity measures which compare density maps [52]. We propose match-distance-based metric to compare the predicted and ground truth pedestrian point sets. Intuitively, we find an optimal matching between the two point sets, and the matching distance is used as the similarity measure. For each unmatched point, we add a constant penalty. The penalty constant should be proportional to the image patch size, essentially the worst possible matching distance between any two points. See Fig. 2(a) for an illustration.

Formally, the proposed similarity between two given pedestrian point sets, $P = \{p_i \in \mathbb{R}^2 \mid i = 1, \ldots, N_P\}$ and $Q = \{q_i \in \mathbb{R}^2 \mid i = 1, \ldots, N_Q\}$. Without loss of generality, we assume $N_P \leq N_Q$. We define the spatial matching distance between them as

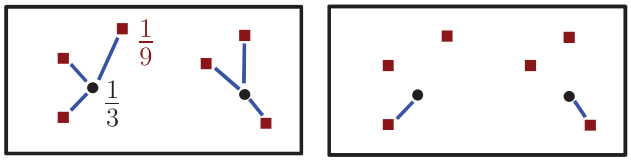
$$\text{dist}(P, Q) = \frac{M(P, Q) + H(P, Q)}{N_Q}, \quad (2)$$

in which $M(P, Q)$ and $H(P, Q)$ are the matching distance and the penalty for unmatched points in Q. Formally,

$$M(P, Q) = \min_{\gamma \in \Gamma(P,Q)} \sum\nolimits_{p \in P} \|p - \gamma(p)\|_2 \quad (3)$$

$$H(P, Q) = (N_Q - N_P) \cdot C, \quad (4)$$

in which Γ is the set of all possible one-to-one mappings from P to Q. We choose the penalty C for each unmatched point in Q to be the diagonal length of the image patch, i.e., $C = \sqrt{\text{height}^2 + \text{width}^2}$. In practice, for an image patch, I, given a prediction pedestrian point set $Pred(I)$ and a ground truth pedestrian point set $GT(I)$, we simply pick the one with the smaller cardinality as P, and the other as Q. For convenience, we will abuse the notation and denote the distance as $\text{dist}(Pred(I), GT(I))$.



(a) Discrete Wasserstein (b) Ours & Hungarian

Figure 3: Distance between prediction (red) and GT (black). Discrete Wasserstein gives a fractal mass to each point and matches all of them (blue lines). Our method only conducts one-to-one matching. Unmatched points get a large penalty. Hungarian loss drops unmatched points.

Due to the stochastic nature of deep learning optimizer, the prediction on a single epoch may not be reliable. We stabilize the proposed spatial matching distance by accumulating over snapshot models at different training epochs. Formally, our accumulated spatial matching distance (ASM) for a training patch, I, is as follows:

$$ASM(I) = \frac{1}{T} \sum_{t=1}^{T} \text{dist}(Pred_t(I), GT(I)) \quad (5)$$

where $Pred_t(I)$ is the model predictions at the t-th training epoch. T is the total number of training epochs. Intuitively, a patch has a lower ASM if it has a better prediction more frequently. A better prediction means the predicted pedestrian point set is better matched with the ground truth points. As shown in Fig. 2 (b) and (c), our ASM can be a fairly good surrogate for uncertainty estimation.

Difference between Hungarian loss [49] & discrete Wasserstein distance. The key difference is two-fold. First, our method is the very first to use such loss for the purpose of uncertainty estimation. We have empirically established how powerful it is in semi-supervised crowd counting tasks. Second, from a technical point of view, these methods are not penalizing unmatched points while our method does. The Hungarian loss drops unmatched points as false proposals. The discrete Wasserstein distance treats points as masses and matches them despite a cardinality discrepancy (see Fig. 3). Therefore, these methods do not fit our task; we expect a prediction with significantly fewer/more points than GT to have a large uncertainty/penalty.

Why patch uncertainty is better than pixel uncertainty? Pixel-wise uncertainty only uses local information. It is similar to an image segmentation uncertainty map, in which only pixels near object boundaries have high uncertainty. Similarly, as illustrated in Fig. 5 (Pixel Uncertainty column), the pixel-wise uncertainty map has high uncertainty

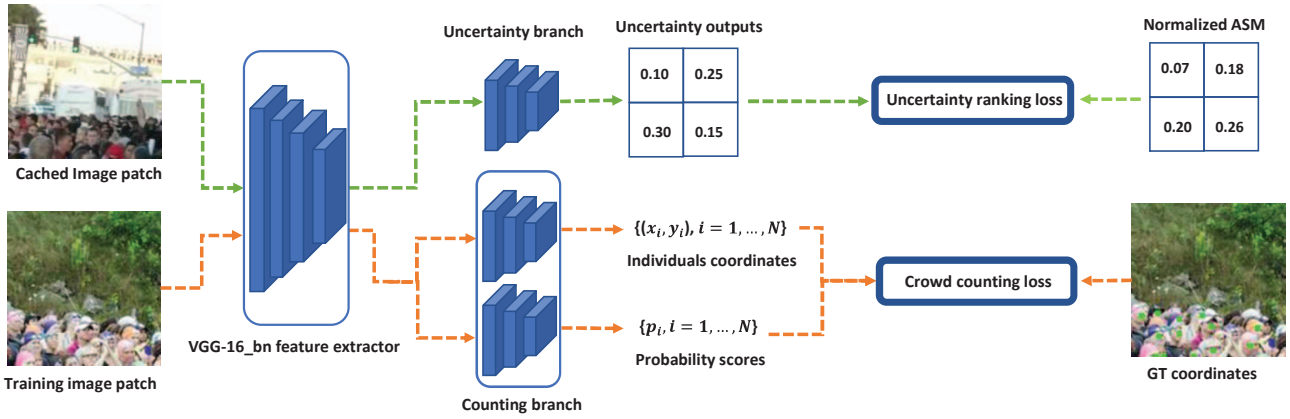


Figure 4: The pipeline of model training for crowd counting and uncertainty estimation. The cached image goes through the feature extractor and uncertainty branch for generating uncertainty scores. These uncertainty scores are supervised by ASM using uncertainty ranking loss. The training image patches are cropped from training images at random locations, and their extracted features are fed into the counting branch for generating the coordinates and probability scores of individuals, which are supervised by ground truth through crowd counting loss.

between dense and sparse regions. But inside a dense region or inside a sparse region, the uncertainty is low despite the prediction quality. On the other hand, our method uses a patch-wise matching-based surrogate function for uncertainty. It is reliable even inside dense or sparse regions.

Implementation details: patch bank. It is a common practice in crowd counting to extract patches at random locations during training. These random patches are different across different training epochs. Thus we cannot accumulate their matching distances. To address this issue, we choose a fixed patch extraction strategy. We cut input images into a fixed set of patches, and store all these patches into a patch bank. During training, for every patch in the bank, we accumulate their matching distance as a surrogate measure for uncertainty. These patches are randomly drawn to train the model within each training batch. As shown in Fig. 4, a cached image patch with size 128×128 is fed into a model for generating 4 uncertainty outputs. Each output represents the uncertainty of the corresponding 64×64 patch region. The uncertainty outputs are supervised by normalized ASM through uncertainty ranking loss. Training image patches, cropped from training images randomly, are used to train the model for crowd counting.

3.3. Semi-Supervised Crowd Counting

During training, our method starts with a labeled data set $\mathcal{D}_L$ and an unlabeled data set $\mathcal{D}_U$. As the training continues, unlabeled data with low uncertainty are assigned pseudo-labels, constituting a pseudo-labeled data set, $\mathcal{D}_{Pseudo}$.

We use two models: a teacher model, f_{tea}, and a student model, f_{stu}. The teacher model makes predictions on unlabeled data set. It generates predictions and estimates uncertainties for all unlabeled patches. Predictions with low uncertainty are chosen as pseudo-labels. These labels, together with their corresponding patches, are added to $\mathcal{D}_{Pseudo}$. Both $\mathcal{D}_L$ and $\mathcal{D}_{Pseudo}$ are used for training. Meanwhile, the student model continuously learns from the training set to make predictions and estimate uncertainty. To train the student model to predict, we use a standard crowd counting loss, $\mathcal{L}_{pred}$, which is backbone dependent. To train the student model to estimate uncertainty, we use the uncertainty loss $\mathcal{L}_{uncer}$, as defined in Eq. 1. Altogether, the student model is trained with the loss

$$\mathcal{L}(f_{stu}, \mathcal{D}) = \mathcal{L}_{pred}(f_{stu}, \mathcal{D}_L) + \mathcal{L}_{uncer}(f_{stu}, \mathcal{D}_L) + \lambda_1 \mathcal{L}_{pred}(f_{stu}, \mathcal{D}_{Pseudo}), \quad (6)$$

in which the weight λ_1 is tuned empirically. Note that while both labeled and pseudo-labeled data are used to train prediction, we only use the original labeled data to train uncertainty, in order to ensure a reliable uncertainty estimation.

We use the same backbone for both the teacher and the student models. In particular, we use P2PNet [48] and its associated prediction loss, $\mathcal{L}_{pred}$. Both models are initialized with the weights pre-trained on ImageNet. We follow the common practice in semi-supervised learning [47, 50] and update the teacher model based on the student model through the exponential mean average (EMA) strategy. With this strategy, the teacher network is more stable and reliable. We depend on its prediction and uncertainty estimation to generate pseudo-labels.

We conclude this subsection by explaining our strategy of pseudo-label generation. At every training epoch, we apply the teacher model to all unlabeled data, generating predictions and uncertainties. For each data (image patch), we compare its uncertainty with a preset uncertainty threshold,

u_t. We select all patches with $< u_t$ uncertainty and their predictions as the pseudo-labeled set. See Fig. 1 for illustrations. Empirically, we adjust the threshold u_t as training progresses. We start with a warm up period, during which we only train the student model with labeled patches. After the warm up period, we use a linearly increasing threshold u_t to select pseudo-labeled patches and add them to the training set. This ensures the pseudo-label generation to be conservative in the beginning and more aggressive later, when the model is more reliable. When adding pseudo-labeled patches, we also apply strong augmentation, i.e., cutout. More details are provided in the supplementary.

Model architecture details. We use the first 13 convolution layers of model VGG-16_bn [42] as the features extractor. The features are fed into two branches: the counting branch is the same as in [48]. The uncertainty branch comprises of convolution layers with ReLU activations. The final layer uses a sigmoid activation to generate the uncertainty.

4. Experiments

In this paper, we conduct extensive experiments on five public datasets to evaluate the effectiveness of our method: ShanghaiTech Part-A and Part-B [64], UCF-QNRF [10], NWPU-Crowd [53], JHU-Crowd++ [45, 46]. More implementation details can be found in Appendix.

Method	Ratio	Part A		Part B	
		MAE	RMSE	MAE	RMSE
SUA [34]	50%	68.5	121.9	14.1	20.6
GP [44]	25%	91	149	-	-
MT [50]	10%	94.5	156.1	15.6	24.5
L2R [26]	10%	90.3	153.5	15.6	24.4
IRAST [29]	10%	86.9	148.9	14.7	22.9
IRAST+SPN [29]	10%	83.9	140.1	-	-
AL-AC [65]	10%	80.4	138.8	12.7	20.4
PA [59]	10%	<u>72.79</u>	**111.61**	12.03	<u>18.70</u>
DAcount [20]	10%	74.9	<u>115.5</u>	<u>11.1</u>	19.1
Ours	10%	**70.76**	116.62	**9.71**	**17.74**

Table 1: Results on the ShanghaiTech dataset.

Data processing. During training, we randomly scale input images with scaling range $[0.7, 1.3]$ and crop patches of size 128×128. The cropped patches are randomly flipped with a probability of 0.5. For some datasets with very high resolution images, e.g., QNRF-Crowd, JHU-Crowd++ and NWPU-Crowd, we rescale the images so the max sizes of images are shorter than a certain length. Following the settings in P2PNet [48], for QNRF-Crowd, JHU-Crowd++, and NWPU-Crowd, this length is 1408, 1430, and 1920.

Evaluation metrics. We use two very common metrics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), to evaluate the model performance.

Hyperparameters. Here we estimate the uncertainty for 64×64 image patches. Thus the penalty constant is $C = 64\sqrt{2}$. Our method is trained by Adam [14] with a mini-batch size of 8 using a learning rate of 1e-5 for the parameters of the feature extractor and 1e-4 for the rest model parameters. The weight λ_1 is set as 0.3. The uncertainty surrogates for images in the bank are updated every training cycle on unlabeled images.

Baselines. To show how our proposed semi-supervised approach can better utilize unlabeled images and boost performance on crowd counting, we compare its performance against SOTA methods from three tracks: semi-supervised learning (SSL), active learning (ACL), and partial-supervised learning (PAL).

4.1. Results

In this part, we compare our semi-supervised method, trained with only 10% of the ground truth labels, to various baselines and show the superiority of our method through evaluation metrics MAE and RMSE. More experiment results can be found in Appendix.

ShanghaiTech. The ShanghaiTech dataset is constructed by two subsets: Part-A and Part-B. The images of Part-A contain dense crowds collected from the Internet. The Part-B images are acquired from a street in Shanghai, which has relatively sparse crowds. Our method achieves superior performance on both Part-A and Part-B, especially in the MAE metric. Tab. 1 indicates that our method can handle both congested and sparse crowds properly and utilize the information from unlabeled images in a good manner.

UCF-QNRF. UCF-QNRF is a challenging crowd counting dataset due to the diversity of viewpoints, crowd densities, and lighting conditions. Our method has a very good performance on this dataset. The results in Tab. 2 indicate that our method is robust to the distribution of the complex background and can generate reliable pseudo-labels under multiple scene situations, achieving lower MAE and RMSE over SOTA semi-supervised methods.

JHU-Crowd++. JHU-Crowd++ is a large-scale crowd counting dataset containing images with weather-based degradations and multiple environmental conditions. Our method achieves better results than all baselines on both MAE and RMSE with DACount [20] at a close second. The results shown in Tab. 3 reflect that our method can utilize the crowd information in unlabeled images efficiently under various scenarios.

NWPU-Crowd. NWPU-Crowd is a massive crowd counting dataset containing images with highly congested crowds and large appearance variations. As shown in Tab. 4, our method has the best results among weakly supervised algorithms. Besides, our method reduces MAE by 39.1%

Method	Type	Ratio	MAE	RMSE
SUA [34]	SSL	50%	130.3	226.3
GP [44]	SSL	25%	147	226
IRAST [29]	SSL	20%	135.6	233.4
MT [50]	SSL	10%	145.5	250.3
L2R [26]	SSL	10%	148.9	249.8
PA [59]	PAL	10%	128.13	218.05
DACount [20]	SSL	10%	109.0	187.2
Ours	SSL	10%	**104.04**	**164.25**

Table 2: Results on the UCF-QNRF.

Method	Type	Ratio	MAE	RMSE
SUA [34]	SSL	50%	80.7	290.8
MT [50]	SSL	10%	90.2	319.3
L2R [26]	SSL	10%	87.5	315.3
PA [59]	PAL	10%	129.65	400.47
DACount [20]	SSL	10%	75.9	282.3
Ours	SSL	10%	**74.87**	**281.69**

Table 3: Results on JHU-Crowd++.

Method	Type	Ratio	MAE	RMSE
MT [50]	SSL	50%	129.8	515.0
L2R [26]	SSL	50%	125.0	501.9
SUA [34]	SSL	50%	111.7	443.2
PA [59]	PA	10%	178.70	1080.43
Ours	SSL	10%	**108.78**	**458.02**

Table 4: Results on NWPU-Crowd.

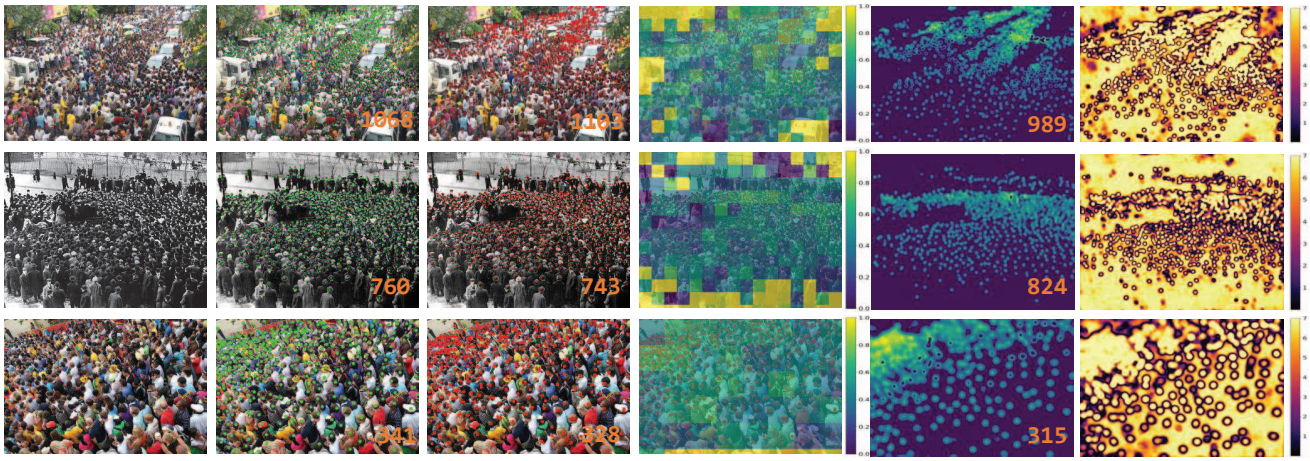


Figure 5: Qualitative results on ShanghaiTech Part-A. The "Prediction" column shows our prediction results. The "Uncertainty" column shows our patch-wise uncertainty, in which yellow represents low uncertainty. The "Density" column is the density map generated by Meng et al. [34] and the "Pixel uncertainty" column shows their pixel uncertainty.

and RMSE by 57.6% compared with the partial annotation-based method (PA) [59].

The results show that our semi-supervised method consistently outperforms the state-of-the-art semi-supervised, partial-annotation, and active learning methods. This reflects that our uncertainty estimation is efficient for filtering out low-quality pseudo-labels. Hence, our semi-supervised framework can utilize the information in unlabeled images more effectively. Sample results are in Fig. 5.

4.2. Ablation Studies

We conduct experiments to illustrate the effectiveness of each component in our method and the effect of changing hyperparameters and experiment settings.

The proportion of labeled images. In the previous section, we conduct experiments with 10% labeled images on ShanghaiTech Part-A. To show the effect of labeled image ratio on the performance of our method, here we use extra experiment results on 5% and 40% labeled images to validate our method. As shown in Tab. 5, our method achieve better performance than baselines despite the ratio of the labeled images. These results indicate our method can utilize unlabeled images efficiently with different proportions of labeled images.

Method	Type	Ratio	Part A	
			MAE	RMSE
MT [50]	SSL	5%	104.7	156.9
L2R [26]	SSL	5%	103.0	155.4
GP	SSL	5%	102.0	172.0
PA [59]	PAL	5%	79.42	**123.60**
DAcount [20]	SSL	5%	85.4	134.5
ours	SSL	5%	**74.48**	127.51
MT [50]	SSL	40%	88.2	151.1
L2R [26]	SSL	40%	86.5	148.2
DAcount [20]	SSL	40%	67.5	110.7
Ours	SSL	40%	**64.74**	**109.56**

Table 5: The ablation study results of labeled ratio on the ShanghaiTech dataset.

Components. Here we use experiments on 10% ShanghaiTech Part-A to verify the effectiveness of components for our method. The results are shown in Tab. 7. Sup.only is the result of the model trained only with supervised counting loss. The result using supervised counting loss and uncer-

tainty loss is shown in Sup.+uncer. There is little improvement in counting performance using our uncertainty loss. To verify the effectiveness of Lambdaloss, here we use L1 loss to supervise the learning process of uncertainty estimation. The results indicate the uncertainty estimation learned with Lambdaloss is more reliable and thus achieves better results. Besides, we also conduct experiments for strong augmentation and EMA. As shown in W/o strong, the lack of strong augmentation has a negative effect on the model performance. The results of W/o EMA indicate EMA is important for utilizing pseudo-labels during the training. An ablation study shows that Hungarian loss has worse performance than our spatial matching distance (Hungarian in Tab. 7).

Method	Ratio	Part A	
		MAE	RMSE
W/o filtering	10%	83.28	172.97
Softmax	10%	75.47	129.35
ACD	10%	71.90	122.03
AWD	10%	72.47	126.12
w/o average	10%	71.47	120.07
Ours	10%	**70.76**	**116.62**

Table 6: The ablation study results of Uncertainty estimation on the ShanghaiTech Part-A.

Method	Ratio	Part A	
		MAE	RMSE
Sup.only	10%	77.74	125.86
Sup.+uncer.	10%	74.84	122.95
L1 loss	10%	72.60	117.72
W/o strong	10%	72.77	124.64
W/o EMA	10%	75.38	124.48
Hungarian	10%	72.95	121.24
ours	10%	**70.76**	**116.62**

Table 7: The ablation study results of components on the ShanghaiTech Part-A.

Uncertainty estimation. We study the effect of uncertainty estimation on our semi-supervised method. Here we show the superiority of our uncertainty estimation method through experiments on 10% ShanghaiTech Part-A. We have three baselines for uncertainty estimation: w/o filtering, softmax, accumulated counting difference (ACD), and accumulated Wasserstein distance (AWD). W/o filtering here represents the baseline without filtering out high uncertainty patches. Softmax is the patch uncertainty estimation calculated by the mean confidence scores of the point proposals in each image patch. ACD is the uncertainty surrogate defined by substituting our spatial matching distance with the absolute counting difference between ground truth and prediction. AWD uses discrete Wasserstein distance instead of spatial matching distance to measure the localization difference. To deal with the case when prediction (ground truth) counting is zero and ground truth (prediction) counting is non-zero for AWD, we apply punishment on such case by constructing a super-pixel, the distance of which to all ground truth points is penalty constant C. As shown in Tab. 6, our method achieves the best performance among all baselines, which reflects that our uncertainty estimation is reliable for choosing high-quality pseudo-labels. From the result of W/o filtering, we know the noisy pseudo-labels are detrimental to the training process and can lead to inferior results. Due to the severe overconfidence problem, the pseudo-labels generated with softmax are still noisy. ACD ignores the location information of individuals in crowds, which is critical for estimating model uncertainty. As discussed in Sec. 3.2, limited by the drawback of discrete Wasserstein distance in evaluating point distribution difference for crowd counting, the performance of AWD is thus worse than our method. In Tab. 7 w/o average, we show the necessity of accumulating sptial matching distance during training.

Hyperparameters. In Tab. 8, we study the effect of the weight λ_1 in Eq. 6, the maximum value of uncertainty threshold u_t, and patch size for uncertainty estimation. We can see our method is not sensitive to those hyperparameters. Our method achieves fairly good performance with the perturbation of hyperparameters.

Patch size	λ_1	max u_t	Ratio	MAE	RMSE
8	0.3	0.6	10%	76.31	126.19
16	0.3	0.6	10%	77.47	133.69
32	0.3	0.6	10%	71.37	116.74
64	**0.3**	**0.6**	10%	**70.76**	**116.62**
128	0.3	0.6	10%	75.41	119.94
64	0.1	0.6	10%	72.73	120.18
64	0.2	0.6	10%	71.79	119.50
64	**0.3**	**0.6**	10%	**70.76**	**116.62**
64	0.4	0.6	10%	72.09	124.64
64	0.5	0.6	10%	70.98	122.00
64	0.3	0.8	10%	73.15	127.48
64	0.3	0.7	10%	71.82	120.23
64	**0.3**	**0.6**	10%	**70.76**	**116.62**
64	0.3	0.5	10%	73.11	122.18
64	0.3	0.4	10%	72.87	124.39

Table 8: The ablation study results of hyperparameters on the ShanghaiTech Part-A.

5. Conclusion

In this work, we introduce a novel patch-wise uncertainty estimation for pseudo-labeling-based semi-supervised crowd counting. Our method trains the uncertainty estimator directly through a surrogate function calculated on labeled patches. As for the uncertainty surrogate function, we use a spatial matching distance between predictions and ground truth labels. Our method provides reliable uncertainty estimation, thus helping to select pseudo-labels to improve the model training in a semi-supervised fashion. The evaluation of our proposed semi-supervised method on several popular crowd counting benchmarks shows that our method consistently achieves superior performance compared to SOTA semi-supervised methods.

Acknowledgements. We thank the anonymous reviewers for their constructive feedback. This work was supported by the NSF grant CCF-2144901.

References

[1] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. In *AAAI*, 2021.

[2] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *CVPR*, 2020.

[3] 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 *MICCAI*, 2019.

[4] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *BMVC*, 2020.

[5] Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *ICLR*, 2018.

[6] Saumya Gupta, Yikai Zhang, Xiaoling Hu, Prateek Prasanna, and Chao Chen. Topology-aware uncertainty for image segmentation. *arXiv preprint arXiv:2306.05671*, 2023.

[7] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *CVPR*, 2019.

[8] Xiaoling Hu, Dimitris Samaras, and Chao Chen. Learning probabilistic topological representations using discrete morse theory. In *ICLR*, 2023.

[9] Yutao Hu, Xiaolong Jiang, Xuhui Liu, Baochang Zhang, Jungong Han, Xianbin Cao, and David Doermann. Nascount: Counting-by-density with neural architecture search. In *ECCV*, 2020.

[10] 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 *ECCV*, 2018.

[11] 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 *CVPR*, 2019.

[12] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In *CVPR*, 2020.

[13] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[15] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *NeurIPS*, 2018.

[16] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Where are the blobs: Counting by localization with point supervision. In *ECCV*, 2018.

[17] Chen Li, Xiaoling Hu, and Chao Chen. Confidence estimation using unlabeled data. In *ICLR*, 2023.

[18] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, 2018.

[19] 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 *CVPR*, 2019.

[20] Hui Lin, Zhiheng Ma, Xiaopeng Hong, Yaowei Wang, and Zhou Su. Semi-supervised crowd counting via density agency. In *ACM MM*, 2022.

[21] Chenchen Liu, Xinyu Weng, and Yadong Mu. Recurrent attentive zooming for joint crowd counting and precise localization. In *CVPR*, 2019.

[22] Liang Liu, Hao Lu, Haipeng Xiong, Ke Xian, Zhiguo Cao, and Chunhua Shen. Counting objects by blockwise classification. *TCSVT*, 2019.

[23] Liang Liu, Hao Lu, Hongwei Zou, Haipeng Xiong, Zhiguo Cao, and Chunhua Shen. Weighing counts: Sequential crowd counting by reinforcement learning. In *ECCV*, 2020.

[24] N Liu, Y Long, C Zou, Q Niu, L Pan, H Adcrowdnet Wu, et al. An attention-injective deformable convolutional network for crowd understanding. In *CVPR*, 2019.

[25] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *CVPR*, 2019.

[26] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *CVPR*, 2018.

[27] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Exploiting unlabeled data in cnns by self-supervised learning to rank. *TPAMI*, 2019.

[28] Xiyang Liu, Jie Yang, Wenrui Ding, Tieqiang Wang, Zhijin Wang, and Junjun Xiong. Adaptive mixture regression network with local counting map for crowd counting. In *ECCV*, 2020.

[29] Yan Liu, Lingqiao Liu, Peng Wang, Pingping Zhang, and Yinjie Lei. Semi-supervised crowd counting via self-training on surrogate tasks. In *ECCV*, 2020.

[30] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. Point in, box out: Beyond counting persons in crowds. In *CVPR*, 2019.

[31] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *ICCV*, 2021.

[32] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *ICCV*, 2019.

[33] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *TMI*, 2020.

[34] Yanda Meng, Hongrun Zhang, Yitian Zhao, Xiaoyun Yang, Xuesheng Qian, Xiaowei Huang, and Yalin Zheng. Spatial uncertainty-aware semi-supervised crowd counting. In *ICCV*, 2021.

[35] Yunqi Miao, Zijia Lin, Guiguang Ding, and Jungong Han. Shallow feature based dense attention network for crowd counting. In *AAAI*, 2020.

[36] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *ICML*, 2020.

[37] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *ECCV*, 2018.

[38] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, and R Venkatesh Babu. Locate, size, and count: accurately resolving people in dense crowds via detection. *TPAMI*, 2020.

[39] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *CVPR*, 2018.

[40] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting perspective information for efficient crowd counting. In *CVPR*, 2019.

[41] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *CVPR*, 2018.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[43] Vishwanath A Sindagi and Vishal M Patel. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *ICCV*, 2019.

[44] 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. In *ECCV*, 2020.

[45] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *ICCV*, 2019.

[46] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *Technical Report*, 2020.

[47] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.

[48] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yang Wu. Rethinking counting and localization in crowds: A purely point-based framework. In *ICCV*, 2021.

[49] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *CVPR*, 2016.

[50] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017.

[51] Jia Wan, Wenhan Luo, Baoyuan Wu, Antoni B Chan, and Wei Liu. Residual regression with semantic prior for crowd counting. In *CVPR*, 2019.

[52] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. In *NeurIPS*, 2020.

[53] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *TPAMI*, 2020.

[54] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *CVPR*, 2019.

[55] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Pixel-wise crowd understanding via synthetic data. *IJCV*, 2021.

[56] Xuanhui Wang, Cheng Li, Nadav Golbandi, Michael Bendersky, and Marc Najork. The lambdaloss framework for ranking metric optimization. In *CIKM*, 2018.

[57] 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 *ICCV*, 2019.

[58] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *ICCV*, 2021.

[59] Yanyu Xu, Ziming Zhong, Dongze Lian, Jing Li, Zhengxin Li, Xinxing Xu, and Shenghua Gao. Crowd counting with partial annotations in an image. In *ICCV*, 2021.

[60] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Reverse perspective network for perspective-aware object counting. In *CVPR*, 2020.

[61] Yifan Yang, Guorong Li, Zhe Wu, Li Su, Qingming Huang, and Nicu Sebe. Weakly-supervised crowd counting learns from sorting rather than locations. In *ECCV*, 2020.

[62] 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 *MICCAI*, 2019.

[63] Qi Zhang and Antoni B Chan. Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In *CVPR*, 2019.

[64] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, 2016.

[65] Zhen Zhao, Miaojing Shi, Xiaoxiao Zhao, and Li Li. Active crowd counting with limited supervision. In *ECCV*, 2020.