Optimal Transport Minimization: Crowd Localization on Density Maps for Semi-Supervised Counting

Wei Lin  Antoni B. Chan
Department of Computer Science, City University of Hong Kong
elonlin24@gmail.com, abchan@cityu.edu.hk

Abstract

The accuracy of crowd counting in images has improved greatly in recent years due to the development of deep neural networks for predicting crowd density maps. However, most methods do not further explore the ability to localize people in the density map, with those few works adopting simple methods, like finding the local peaks in the density map. In this paper, we propose the optimal transport minimization (OT-M) algorithm for crowd localization with density maps. The objective of OT-M is to find a target point map that has the minimal Sinkhorn distance with the input density map, and we propose an iterative algorithm to compute the solution. We then apply OT-M to generate hard pseudo-labels (point maps) for semi-supervised counting, rather than the soft pseudo-labels (density maps) used in previous methods. Our hard pseudo-labels provide stronger supervision, and also enable the use of recent density-to-point loss functions for training. We also propose a confidence weighting strategy to give higher weight to the more reliable unlabeled data. Extensive experiments show that our methods achieve outstanding performance on both crowd localization and semi-supervised counting. Code is available at https://github.com/Elin24/OT-M.

1. Introduction

Crowd understanding gains much attention due to its wide applications in surveillance [33, 61] and crowd disaster prevention. Most studies in this area concentrate on crowd counting, whose objective is to provide the total number and distribution of crowds in a scene automatically. Due to the development of deep learning, recent methods [5, 58, 59, 62, 67] have achieved success on a variety of counting benchmarks [50, 61, 62, 66]. Counting methods can be extended to other applications, such as traffic management [63], animal protection [2], and health care [32].

Although crowd counting has been greatly developed, most methods do not explore further applications of the estimated density maps after obtaining the count. Specifically, there is limited research on crowd localization, tracking, or group analysis with predicted density maps. Taking crowd localization as an example, only a few methods, such as local maximum [13, 58, 65], integer programming [36], or Gaussian mixture models (GMM) [14], have been proposed to locate pedestrians or tiny objects from density maps. Moreover, recent localization methods ignore the counting density map, and instead are based on point detection [40, 54], blob segmentation [1, 10, 17], or inverse distance maps [24]. However, this will increase the inefficiency of a crowd understanding system, since separate networks are required for counting and localization.

To broaden the application of density maps for localization, in this paper we propose a parameter-free algorithm, Optimal Transport Minimization (OT-M), to estimate the point map indicating locations of objects from a counting density map (see Fig. 1). OT-M minimizes the Sinkhorn distance [7] between the density map (source) and point localization map (target), through an alternating scheme that estimates the optimal transport plan from the current point map (the OT-step), and updates the point map by minimizing their transport cost (the M-step). OT-M is parameter-free and requires no training, and thus can be applied to any crowd-counting method using density maps.

To demonstrate the applicability of density-map based localization, we apply OT-M to semi-supervised counting. In previous work, [38] builds a baseline for semi-supervised counting based on the mean-teacher framework [55], but finds it ineffective. Looking closely, we note that the baseline in [38] uses a soft pseudo-label (density map) to supervise the student model, whereas successful semi-supervised classification [52] or segmentation [6, 55] methods usually
are based on hard pseudo-labels (e.g., class labels or binary segmentation masks). In the context of semi-supervised crowd counting, the hard label is the point map in which each person is pseudo-annotated with a point. Thus, in this paper we generate hard pseudo-labels using OT-M for the unlabeled crowd images for semi-supervised crowd counting. As an additional benefit, the hard pseudo-labels allow training CNNs under semi-supervised learning using recent density-to-point loss functions (e.g., Bayesian loss (BL) [34] or generalized loss (GL) [58]), which are more effective than traditional losses, e.g. L2.

Similar to other semi-supervised tasks, some estimated pseudo labels may be inaccurate due to limitations of the current trained model. To reduce the effect of these noisy pseudo-labels, we propose a confidence-weighted GL (C-GL) for semi-supervised counting. Specifically, we compute the unbalanced optimal transport plan between the student’s predicted density map (source) and the hard pseudo-labels from the teacher (target), and then define the pixel-wise and point-wise confidences based on the consistencies between source, target and the plan. Experiments show that the trained model is more robust with our C-GL.

In summary, the contributions of this paper are 3-fold:

• We propose an OT-M algorithm to estimate the locations of objects from density maps, which is based on minimizing the Sinkhorn distance between the density map and the target point map. Since OT-M is parameter-free, it can be applied to any density map without training.

• We use OT-M to produce hard pseudo-labels for semi-supervised counting, which conforms with schemes in other semi-supervised tasks. The hard label also allows applying density-to-point loss like GL to unlabeled data for more effective training.

• To mitigate risks brought by inaccurate pseudo-labels, we propose a confidence-weighted Generalized Loss to reduce the influence of inconsistencies between the teacher’s and student’s predictions. Experiments show that our loss improves semi-supervised counting performance.

2. Related works

Crowd Counting. Before deep learning became popular, detection-based methods for counting pedestrians were based on detecting human body parts [19, 20], but do not work well in the dense crowds due to partial occlusions. Instead, regression-based methods overcome these obstacles by directly predicting the final count based on low-level features [4, 5, 12]. Recent methods use a convolutional neural network (CNN) to estimate a density map from a crowd image, and the corresponding count is obtained by summing over the density map [21]. Various networks are proposed to address scale variations in crowd scene [3, 11, 22, 67], by using multiple columns [3, 67], multi-task learning [11], or dilated convolution [22]. However, all these methods are trained by pixel-wise L2 loss, which is ineffective since the original ground-truth point map is blurred by a hand-crafted Gaussian kernel and loses localization information. To address this, [25, 34, 35, 58, 60] directly compare the predicted density map and the ground-truth point map. [35] designed an efficient algorithm to optimize counting models based on UOT’s semi-dual regularized formulation, while [25] derived a semi-balanced form of Sinkhorn divergence to satisfy the identity of indiscernibles. [58] proves that L2 loss and Bayesian loss [34] are special cases of generalized loss.

Localization with Object Density Maps. The goal of most crowd counting methods is to estimate a density map representing the distribution of pedestrians, and then take its sum as the final count result. However, as shown in Fig. 1, localizing pedestrians according to the density map is tricky since the predicted density map is blurry. [36] recovers the locations of objects by applying integer programming to windowed observations of the density map. [36] also considers clustering methods, like K-means or mean-shift, on each connected component to localize partially-occluded instances. Differently, [14] proposes localization by learning a Gaussian mixture model (GMM) to fit the density map, where the centers of the estimated Gaussian components correspond to the people locations. Finally, [13, 58] aim to estimate sparse density maps, and then define every local maximum pixel whose value is greater than a threshold as the location of a person. In contrast to these methods, our OT-M is based finding the point map with minimal Sinkhorn distance to the density map. Empirical results show that OT-M is more accurate and robust.

Density maps are also used to improve detectors and trackers in crowded scenes [14]. In [45], density estimation is jointly optimized with standard detectors to reduce false positive errors and improve recall. In [43, 44], the tracking-by-counting paradigm is proposed to overcome occlusions and appearance variations during tracking in crowd scene.

Semi-supervised Counting. As labeling very dense crowd images can be expensive, leveraging unlabeled crowd images with semi-supervised counting has seen increased interest in recent years. L2R [29, 30] introduces a rank rule for unlabeled data, inspired by the observation that cropped image contains the same or fewer objects than the original image. GP [49] proposes an iterative method based on the Gaussian process to assign soft pseudo-labels to unlabeled images. IRAST [31] uses density segmentation as a surrogate task to detect conflicting predictions and correct them. The segmentation task is also considered by SUA [38], which follows the popular teacher-student scheme [16], and segmentation results are used to model unlabeled data’s uncertainty. DACount [26] designs a structure similar to the multiple columns and switching mod-
ule in Switch-CNN [3], and uses multiple learnable density agents to learn features of crowd in different density levels.

Unlike these methods based on surrogate tasks, our semi-supervised framework is based on density maps predicted from the unlabeled data. Our OT-M is utilized to generate hard pseudo-labels (point maps) for training with unlabeled data, and we further propose a confidence-weighted GL to reduce the influence of inaccurate pseudo-labels.

3. OT-M Algorithm

In object counting, the ground-truth (GT) density map is obtained by convolving a Gaussian kernel with the GT point map, i.e., converting the hard-label into a soft-label. However, there is little research to address the inverse problem, converting the soft-label density map into a hard-label point map. In this section, we introduce a parameter-free algorithm to estimate the hard label from a soft density map by minimizing the entropic optimal transport cost (i.e., Sinkhorn distance [7]) between them.

Let the soft-label density map predicted by a CNN be represented as \( A = \{ (a_i, x_i) \}_{i=1}^{n} \), where \( a_i \geq 0 \) and \( x_i \in \mathbb{R}^2 \) are the density value and coordinate of the \( i \)-th pixel, and \( n \) is the number of pixels. Given the density map \( A \), our goal is to estimate a hard label \( B = \{ (b_j, y_j) \}_{j=1}^{m} \), where \( b_j = 1 \) and \( y_j \in \mathbb{R}^2 \) represents the \( j \)-th point (person location). The number of points \( m \) is the count obtained from the density map \( A \), rounded to the nearest integer, i.e., \( m = \lceil \sum_{i=1}^{n} a_i \rceil \). Since \( b_j = 1 \) is fixed for all points, we will use the shorthand \( B = \{ y_j \}_{j=1}^{m} \) to reduce clutter.

We estimate the hard labels by minimizing the Sinkhorn distance between the points \( B \) and the density map \( A \),

\[
\hat{B} = \arg\min_{B=\{y_j\}_{j=1}^{m}} \mathcal{L}^\varepsilon(A, B),
\]

(1)

where \( \mathcal{L}^\varepsilon(A, B) \) is the Sinkhorn distance between \( A \) and \( B \), and \( \varepsilon \) is a near-zero weight for the entropic term:

\[
\mathcal{L}^\varepsilon(A, B) = \min_{P \in U(a, b)} \langle C, P \rangle - \varepsilon \mathcal{H}(P),
\]

(2)

\[
= \sum_{i,j} C_{ij} P_{ij} + \varepsilon \sum_{i,j} P_{ij} \log(P_{ij})
\]

(3)

where \( C = [C_{ij}] \) is the cost matrix, \( P = [P_{ij}] \) is the transport plan, and \( \mathcal{H}(P) = -\sum_{i,j} P_{ij} \log(P_{ij}) \) is the entropy of \( P \). Here, the cost matrix element \( C_{ij} = C(x_i, y_j) \) measures the cost when moving a unit mass from \( x_i \) to \( y_j \). We use the squared Euclidean distance as the cost function:

\[
C(x_i, y_j) = \|x_i - y_j\|^2.
\]

(4)

The Sinkhorn distance in (2) finds the optimal transport plan \( P \), whose element \( P_{ij} \) is the mass quantity (i.e., density) transported from \( x_i \) to \( y_j \), that minimizes the total transport cost. In balanced optimal transport, \( P \) is constrained to admissible couplings that preserves the total mass from each \( x_i \) and to each \( y_j \), \( U(a, b) \) def. \( \{ P \in \mathbb{R}^{n \times m} : P1_m = a, P^\top 1_n = b \} \), where \( a = [a_i] \), \( b = [b_j] \), and \( 1_n \) is the vector of \( n \) ones.

To find the solution of (1), we propose the OT-M algorithm that iteratively computes: 1) the optimal transport plan for Sinkhorn distance in (2) while holding the cost matrix fixed; and 2) the optimal cost matrix, parametrized by the points \( B \), while holding the transport plan fixed. Formally, after initialization of the points \( B^{(0)} = \{ y_j^{(0)} \}_{j=1}^{m} \), the \( k \)-th iteration of the OT-M algorithm is:

\[
\text{OT-step: } P^{(k)} = \arg\min_{P \in U(a, b)} \{ C(B^{(k−1)})P \} - \varepsilon \mathcal{H}(P),
\]

(5)

\[
\text{M-step: } B^{(k)} = \arg\min_{B=\{y_j\}_{j=1}^{m}} \langle C(B), P^{(k)} \rangle - \varepsilon \mathcal{H}(P^{(k)}),
\]

(6)

where \( C(B) \) is the cost matrix between \( A \) and \( B \). The details of each step and convergence proof are presented next.

3.1. Optimal Transport Step (OT-Step)

The goal of OT-step is to compute the optimal transport plan \( P^{(k)} \) between \( A \) and \( B^{(k−1)} \). The solution can be formulated as [41]:

\[
P = \exp(u)K\exp(v), \quad K = \exp(-C/\varepsilon),
\]

(7)

where \( u \in \mathbb{R}^n_+ \) and \( v \in \mathbb{R}^m_+ \) are two unknown scaling variables. Then the minimization of (5) can be solved efficiently through the Sinkhorn algorithm – an alternate minimization scheme [46, 51]. Specifically, the following iterations are repeated after initializing \( v \) with an arbitrary positive vector \( v^{(0)} \) (\( v^{(0)} = 1_m \) by default):

\[
u^{(l+1)} = \frac{a}{Kv^{(l)}}, \quad v^{(l+1)} = \frac{b}{K^\top u^{(l+1)}},
\]

(8)

where the division is element-wise.

In summary, the cost matrix \( C^{(k)} = C(B^{(k−1)}) \) is computed from the current points \( B^{(k−1)} \), and the Gibbs kernel matrix \( K^{(k)} = \exp(-C^{(k)}/\varepsilon) \) is calculated. Next, the iterations in (8) are run until convergence, and the transport plan \( P^{(k)} \) is calculated from (7). Note that \( a \) and \( b \) are normalized to make their sums equal to perform balanced OT.

3.2. Minimization step (M-Step)

The M-step computes the new set of points \( B^{(k)} = \{ y_j^{(k)} \} \) by minimizing (6) while keeping \( P^{(k)} \) fixed. Specifically, we rewrite (6) by plugging in the cost function,

\[
B^{(k)} = \arg\min_{\{y_j\}_{j=1}^{m}} \sum_{i,j} P_{ij}^{(k)} C(x_i, y_j),
\]

(9)
and noting that each \( y_j \) can be optimized independently,

\[
y_j^{(k)} = \arg \min_{y_j} \sum_{i=1}^{n} P_{ij}^{(k)} \| x_i - y_j \|^2.
\] (10)

Setting the derivative of (10) equal to zero, the solution is:

\[
\frac{\partial}{\partial y_j} \sum_{i=1}^{n} P_{ij}^{(k)} \| x_i - y_j \|^2 = 0 \Rightarrow y_j^{(k)} = \frac{\sum_{i=1}^{n} P_{ij}^{(k)} x_i}{\sum_{i=1}^{n} P_{ij}^{(k)}}.
\] (11)

In (11), \( y_j^{(k)} \) is the barycenter of masses assigned to \( y_j \) in the transport plan \( P^{(k)} \). The algorithm is summarized in the Supp., and two iterations are visualized in Fig. 2.

### 3.3. Convergence of the OT-M Algorithm

We next prove that our OT-M algorithm converges – after each iteration the estimated \( B^{(k)} \) decreases the Sinkhorn distance in (1) until a local minimum is achieved, at which point it cannot decrease (but will not increase) [37, 39]. Denote the cost matrix in the \( k \)-th iteration as \( C^{(k)} = C(B^{(k-1)}) \). After computing the optimal transport plan \( P^{(k)} \) for cost matrix \( C^{(k)} \) in the OT-step in (5), we have

\[
\langle C^{(k)}, P^{(k)} \rangle - \varepsilon \mathcal{H}(P^{(k)}) \leq \langle C^{(k)}, P^{(k-1)} \rangle - \varepsilon \mathcal{H}(P^{(k-1)}),
\] (12)

since \( P^{(k)} \) is the minimizer over all admissible transport plans. Next, in the M-step in (6), we obtain the optimal \( B^{(k)} \) for fixed transport plan \( P^{(k)} \), and thus

\[
\langle C(B^{(k)}), P^{(k)} \rangle \leq \langle C^{(k)}, P^{(k)} \rangle,
\] (13)

since the cost matrix \( C(B^{(k)}) \) is the minimizer. Noting that \( C(B^{(k)}) = C^{(k+1)} \), we thus obtain

\[
\langle C^{(k+1)}, P^{(k)} \rangle \leq \langle C^{(k)}, P^{(k)} \rangle,
\] (14)

\[
\Rightarrow \langle C^{(k)}, P^{(k-1)} \rangle \leq \langle C^{(k-1)}, P^{(k-1)} \rangle.
\] (15)

Finally, substituting (15) into (12), we obtain the convergence condition:

\[
\langle C^{(k)}, P^{(k)} \rangle - \varepsilon \mathcal{H}(P^{(k)}) \leq \langle C^{(k-1)}, P^{(k-1)} \rangle - \varepsilon \mathcal{H}(P^{(k-1)}),
\] (16)

where the LHS is the Sinkhorn objective for iteration \( k \) and the RHS is for \( k - 1 \). Thus, in each iteration the objective in (1) is non-increasing, and the algorithm converges.

### 4. OT-M Based Semi-Supervised Counting

Using pseudo-labels [18] is an effective method for semi-supervised learning. Most related works on classification [18, 53, 64] and segmentation [6, 9, 68] empirically show that hard labels are more valuable than soft labels. In this section we show how to effectively take advantage of hard pseudo-labels, i.e., point maps, generated through OT-M algorithm for semi-supervised counting. As shown in Fig. 3, we use the mean-teacher framework [55], where an exponential moving average (EMA) is used to update the parameters in the teacher net. For labeled images, the student net is trained with fully-supervised learning on the GT point maps. For unlabeled images, we use the teacher net to generate a soft pseudo-label (density map), and OT-M is applied to produce a hard pseudo-label (point map). Meanwhile, these unlabeled images are perturbed and input into the student net to generate a prediction, which is supervised by the hard pseudo labels. For effective training, we propose a confidence-weighted generalized loss (C-GL) to reduce the effect of inconsistent (noisy) pseudo-labels.

#### 4.1. Generalized Loss with Gating

The Generalized Loss (GL) [58] is based on the unbalanced optimal transport (UOT) problem,

\[
L_{gl}^{c, \tau} = \min_{P \in \mathbb{R}^n \times m} \langle C, P \rangle - \varepsilon \mathcal{H}(P) + \tau D(P, a, b),
\] (17)
where \(a \in \mathbb{R}^n\) and \(b = 1_m\) are the predicted density values and ground truth point values. \(D\) is a divergence penalizing marginal deviation, with \(\tau\) controls the degree of imbalance allowed. In GL, \(D(P, a, b)\) is defined as

\[
D_{gl}(P, a, b) = \|P1_m - a\|_2^2 + \|P^\top 1_n - b\|_1. \quad (18)
\]

There is no efficient algorithm to directly implement (17) with \(D_{gl}\) as the divergence, so [58] firstly approximates the optimal \(P\) by solving UOT with KL divergence (KL-UOT), and then plugs \(\hat{P}\) back into (17) to compute the GL. Specifically, applying the Sinkhorn algorithm [41] to KL-UOT yields both the optimal transport plan \(\hat{P}\) and \((f^*, g^*)\), the gradients of \((a, b)\). Thus the generalized loss is rewritten:

\[
L_{\tau, \gamma}^{\epsilon, \tau} = a^\top f^* + b^\top g^* - \epsilon \mathcal{H}(\hat{P}) + \tau_2\|\hat{P}1_m - a\|_2^2 + \tau_1\|\hat{P}^\top 1_n - b\|_1. \quad (19)
\]

Note that \(\hat{P}\) is a function of both \(a\) and \(b\).

In (19) we introduce separate hyperparameters \((\tau_1, \tau_2)\) on the L1/L2 loss terms to “gate” them to improve training. Let \(m_a, m_b, m_b\) be the sum of \(P, a, b\) respectively, which correspond to the total transported density, the count of the predicted density map, and the GT count. In practice we find that the Sinkhorn algorithm sometimes estimates a \(\hat{P}\) whose sum \(m_{\hat{P}}\) is larger than both \(m_a\) and \(m_b\), which is harmful to training. For example, suppose \(m_a < m_b\), then the predicted count is smaller than the GT count, and we hope to increase \(m_a\) to match \(m_b\). However if \(m_b < m_{\hat{P}}\), then the L1 loss term will encourage \(m_a\) to decrease, which also decreases \(m_a\), but this is in conflict to the goal of increasing \(m_a\). Thus, we can set \(\tau_1 = 0\) to ignore the L1 loss term when \(m_a < m_b < m_{\hat{P}}\). Other cases can be handled analogously, resulting in the following “gating” of the L1/L2 loss terms through setting of \((\tau_1, \tau_2)\).

\[
\tau_1 = \begin{cases} 0, & m_a < m_b < m_{\hat{P}}; \\ 0, & m_{\hat{P}} < m_a < m_b, \\ \tau, & \text{otherwise}. \end{cases} \quad \tau_2 = \begin{cases} 0, & m_b < m_a < m_{\hat{P}}; \\ 0, & m_{\hat{P}} < m_a < m_b, \\ \tau, & \text{otherwise}. \end{cases}
\]

We set confidence parameter \(\tau = 0.1\) following [58].

4.2. Confidence Strategy

Semi-supervised learning has a common drawback: predictions for unlabeled data are usually noisy, which leads to confirmation bias [15] towards these errors and consequently learns defective models. To overcome this issue, we build a confidence strategy based on the consistency between the teacher’s hard label and the student’s prediction for semi-supervised counting model trained with GL.

Assume the density map predicted by the student model is \(a\), and the hard pseudo-label predicted by the teacher model (via OT-M) is \(b\), and the KL-UOT transport plan between them is \(\hat{P}\). We calculate the consistency via the point-wise distance between the transport plan and point target \(b\):

\[
w_1 = \exp[-\gamma (\operatorname{diag}(b)^{-1} |\hat{P}^\top 1_n - b|)], \quad (20)
\]

in which \(\gamma > 0\) is a hyperparameter to decrease the confidence, and \(|\cdot|\) is the element-wise absolute value. Note that \(w_1\) is close to 1 as long as the sum of each column of \(\hat{P}\)’s is close to corresponding element \(b_j = 1\).

Next, we propagate \(w_1\) to the pixels to compute confidence values for elements in \(a\). Specifically, pixel-wise confidence \(w_2\) is a weighted sum of elements in \(w_1\), and the weight is computed by normalizing each row in \(\hat{P}\):

\[
w_2 = \frac{w_1}{1_m} \hat{P}w_1. \quad (21)
\]

Embedding \(w_1\) and \(w_2\) into GL, the final formulation is:

\[
L_{\epsilon, \tau, \gamma}^{\alpha, \tau} = a^\top W_2 f^* + b^\top W_1 g^* - \epsilon \mathcal{H}(\hat{P}) + \tau_2\|W_2(\hat{P}1_m - a)\|_2^2 + \tau_1\|W_1(\hat{P}^\top 1_n - b)\|_1, \quad (22)
\]

where \(W_1 = \operatorname{diag}(w_1)\) and \(W_2 = \operatorname{diag}(w_2)\). Fig. 4 visualizes \(w_1\) and \(w_2\) in a simple example.

Note that the original GL in (19) is a special case of the C-GL in (22). Some counting methods [26, 27, 57] report that there may be annotation noise in labeled data, so (22) can also be applied to labeled data to depress noise if a suitable \(\gamma\) is given. In our experiments, we set \(\gamma = 0.5\) for both labeled and unlabeled data.

5. Experiments

In this section, we conduct experiments to demonstrate the efficacy of our OT-M algorithm and its use in semi-supervised counting. In the first part, we use synthetic and real data to empirically show the OT-M algorithm’s
convergence process. The localization performance is also compared with previous density-map based algorithms. In the second part, we compare our framework with previous semi-supervised counting approaches. After that, the ablation study is conducted to show whether each component of our framework works as expected.

5.1. Experiment setup

Localization experiments are conducted on two public datasets, UCF-QNRF [13] and NWPU-Crowd [61]. Semi-supervised counting is tested on four datasets: ShanghaiTech-A and B (ST-A, ST-B) [67], UCF-QNRF [13], and JHU++ [50]. In each dataset, 5%, 10%, 40% of training samples are selected as labeled data. We follow 2 protocols, the single trial version from [26], and a new version based on averaging over 5 random trials for each percentage.

For the OT-M algorithm, we consider three initialization methods of \( B^{(0)} \): 1) Top-k selects the \( m \) pixels with the largest density values as the initial points; 2) Uniform selects the initial \( m \) points uniformly at random; 3) adaptive initialization normalizes the density map into a probability distribution, from which the \( m \) initial points are sampled. We set the maximum number of iterations in OT-M as 16, and also use the following early stopping criteria:

\[
\frac{1}{m} \sum_{j=1}^{m} \| y_j^{(k)} - y_j^{(k-1)} \|_2 < 1 \quad \text{and} \quad \max_j \| y_j^{(k)} - y_j^{(k-1)} \|_2 < \frac{1}{r},
\]

where \( r \) is the down-sampling ratio of CNN (\( r = \frac{1}{8} \) in our experiments). When the average distance moved of points in \( B^{(k)} \) is smaller than 1 pixel, and the maximum moved distance is smaller than \( 1/r \), then the algorithm stops.

5.2. Experiments on OT-M Convergence

We first show the effectiveness and convergence of OT-M on some examples from ST-A. The accuracy of OT-M relies on the precision of estimated density maps. When the density map is perfect, the estimated point map should be extremely similar to the GT. To show this, we generate synthetic density maps by applying a Gaussian kernel (with variance 8) to the GT point map, and then recover the point maps through OT-M. In Fig. 6(a), we present the Sinkhorn distance objective between the predicted point map and the given density map during the iterations, when using different initializations. Using the adaptive initialization yields faster convergence, compared to other initialization methods. In Fig. 5(a), we visualize the iterative convergence process on an example. To better show the effectiveness of OT-M, we use the worst initialization method, top-k, where most of the initial points are in a small area. After initialization, OT-M is able to gradually move the points closer to the targets, and finally yields a point map that is close to the ground-truth.

We next evaluate how well the recovered point maps compare to the GT point maps. We use the GT density maps from UCF-QNRF, downsample them by 1/8 to make the problem more difficult, and then apply OT-M to recover the point maps. The results are presented in the first row of Table 1. OT-M obtains high precision, recall, and F1 score for each percentage. We find that the OT-M algorithm is robust to initialization – the three different initialization methods yield the same results, although they have different convergence speed.

Finally, we visualize a example on density maps predicted from a CNN in Fig. 5(b). Compared with synthetic density maps, CNN predictions are more ambiguous and noisy. The OT-M algorithm works as expected: decreasing the Sinkhorn distance between the estimated point map and the given density map, as displayed in Fig. 6(b). We further evaluate the localization performance of OT-M on predicted density maps in the next section.

5.3. Localization Performance on Density Maps

In this section, we compare OT-M algorithm with two other localization methods based on density maps: Lo-
Density Map & Localization & Precision & Recall & F-measure & & & & & & & \hline
\text{ground-truth} & \text{Local Max} [58] & 0.842 & 0.838 & 0.840 & & & & & & & \text{rescaled (ours)} & 0.891 & 0.910 & 0.912 & & & & & & & \hline
\text{GMM} [14] & 0.749 & 0.732 & 0.736 & & & & & & & \hline
\text{MAN} [27] & 0.620 & 0.487 & 0.544 & & & & & & & \hline
\text{ChfL} [47] & 0.812 & 0.571 & 0.671 & & & & & & & \hline
\text{cvpr'21} & \text{GL} [58] & 0.782 & 0.748 & 0.765 & & & & & & & \hline
\text{cvpr'22} & \text{GMM} [14] & 0.750 & 0.728 & 0.739 & & & & & & & \hline
\text{cvpr'19} & \text{OT-M (ours)} & 0.804 & 0.783 & 0.793 & & & & & & & \hline
\text{cvpr'15} & \text{GL} [58] & 0.892 & 0.732 & 0.836 & & & & & & & \hline
\text{cvpr'19} & \text{GMM} [14] & 0.749 & 0.732 & 0.736 & & & & & & & \hline
\text{cvpr'19} & \text{OT-M (ours)} & 0.772 & 0.755 & 0.760 & & & & & & & \hline
\text{cvpr'21} & \text{ChfL} [47] & 0.812 & 0.571 & 0.671 & & & & & & & \hline
\text{cvpr'22} & \text{GMM} [14] & 0.755 & 0.740 & 0.747 & & & & & & & \hline
\text{cvpr'22} & \text{OT-M (ours)} & 0.780 & 0.765 & 0.772 & & & & & & & \hline
\hline
Table 1. Comparison of different localization methods on UCF-QNRF using different density maps (ground-truth and predicted). Note that the ground-truth density map is downsampled by 1/8 to match the output size of GL and ChfL.

5.4. Semi-Supervised Counting

We next present the results for semi-supervised counting. Table 3 compares ours with previous methods using the same backbone and experiment protocol (i.e., same set of labeled data) from [26], consisting of one trial for each label percentage (5%, 10%, 40%). The DAC results are reproduced to ensure the same experiment design. It shows that OT-M outperforms most semi-supervised counting approaches, especially when there are fewer labeled data (5% and 10%). When the label percentage is increased to 40%, DAC [26] achieves lower MAE and MSE on UCF-QNRF and JHU++, while SUA [38]’s MAE is the lowest on ST-A. However, our framework has the smallest MSE on all these datasets.

In the above experiments, only one trial is used for each label percentage, which is inadequate because the random selection of labeled data strongly influences the counting performance and stability. To investigate this issue, we test DAC [26] and our method in another experiment with multiple trials, where each trial uses different randomly selected labeled data. Here the averaged MAE/MSE over multiple trials is more representative of the algorithm’s performance, compared to using a single trial, especially when the number of labeled samples is small. The experiment results are presented in Table 4. For 5% and 10% labeled data, our framework outperforms DAC [26] on all four datasets. The average and standard deviation of MAE/MSE are much smaller than DAC. For 40% labeled data, DAC [26] has lower MAEs than ours on three datasets, while our model has lower MSE on all datasets. In summary, the combination of OT-M and the proposed confidence strategy can achieve outstanding performance using a simple mean-teacher framework, especially for smaller percentages of labeled data (5% & 10%).

5.5. Ablation Study on Semi-Supervised Counting

We next conduct ablation studies using the 5% labeled data setting on UCF-QNRF [13].

Confidence-weighted GL on labeled data. The top half of Tab. 5 presents the effect of the ($\tau_1, \tau_2$) gating scheme
and confidence-weights on labeled data. With gating, MAE is reduced from 145.59 to 144.48. The gap is small since most loss is from the transport term, \( (C, P) \), but the MAE and MSE still decrease by 1.11 and 1.94. As mentioned in Sec. 4.2, confidence weights can also be applied to labeled data to suppress annotation noise. Relevant experimental results demonstrate its effectiveness – it helps the counting model achieve better performance (MAE: 138.52).

**Hard labels vs. soft labels.** Next, we compare the performance while unlabeled data is used during training. We compare our framework with the soft pseudo-labels (predicted density maps) using L2 loss. We also design a confidence strategy for L2 loss:

\[
    w' = \exp[-\gamma' (\text{diag}(a_t)^{-1}(a_t - a_s))],
\]

where \( a_t \) and \( a_s \) represent density maps predicted by the teacher and student, and \( \gamma' \) is similar to \( \gamma \) in (20). The results are shown in the bottom half of Tab. 5 – the counting model can predict more accurately under the guidance of confidence strategy during training, regardless of using soft or hard labels. However, using hard pseudo-labels for unlabeled samples reduces estimation errors dramatically compared to soft labels. Using the confidence-weights with GL also greatly improves the MAE and MSE (e.g., MAE 125.3 drops to 120.13).

**Localization method.** Finally, we consider different density-map localization methods for generating hard pseudo-labels in our semi-supervised counting framework, as presented in Tab. 6. LM [58] performs even worse than training with only labeled data, which is because the number of points generated by LM could be different from the count in the teacher’s density map, \( i.e., \) the number of local maxima in the density map is not guaranteed to be the sum of the density map. In contrast, both GMM [14] and OT-M

<table>
<thead>
<tr>
<th>Label Percentage</th>
<th>Methods</th>
<th>ST-A MAE</th>
<th>ST-A MSE</th>
<th>ST-B MAE</th>
<th>ST-B MSE</th>
<th>UCF-QNRF MAE</th>
<th>UCF-QNRF MSE</th>
<th>JHU++ MAE</th>
<th>JHU++ MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>DAC [26]</td>
<td>92.9±3.4</td>
<td>148.6±10.3</td>
<td>13.4±2.2</td>
<td>24.6±6.7</td>
<td>122.7±7.8</td>
<td>218.9±14.0</td>
<td>81.2±2.4</td>
<td>313.7±12.2</td>
</tr>
<tr>
<td></td>
<td>OT-M (ours)</td>
<td>86.0±2.2</td>
<td>132.7±3.3</td>
<td>12.8±1.4</td>
<td>22.0±4.5</td>
<td>120.1±7.3</td>
<td>208.9±11.7</td>
<td>80.9±3.1</td>
<td>303.1±9.5</td>
</tr>
<tr>
<td>10%</td>
<td>DAC [26]</td>
<td>84.8±4.5</td>
<td>140.9±11.3</td>
<td>11.1±0.5</td>
<td>18.9±1.9</td>
<td>110.3±5.9</td>
<td>196.0±16.3</td>
<td>76.0±2.0</td>
<td>293.8±10.4</td>
</tr>
<tr>
<td></td>
<td>OT-M (ours)</td>
<td>81.6±2.6</td>
<td>127.1±3.8</td>
<td>10.9±0.5</td>
<td>18.1±4.1</td>
<td>107.9±4.1</td>
<td>180.6±7.8</td>
<td>75.5±1.6</td>
<td>287.9±11.1</td>
</tr>
<tr>
<td>40%</td>
<td>DAC [26]</td>
<td>71.6±2.0</td>
<td>120.8±5.6</td>
<td>9.0±0.3</td>
<td>14.6±0.5</td>
<td>91.8±4.7</td>
<td>161.4±12.4</td>
<td>64.1±3.0</td>
<td>270.6±9.3</td>
</tr>
<tr>
<td></td>
<td>OT-M (ours)</td>
<td>70.0±2.2</td>
<td>113.0±6.9</td>
<td>9.0±0.4</td>
<td>14.2±0.7</td>
<td>93.4±5.4</td>
<td>157.5±7.8</td>
<td>66.5±3.1</td>
<td>268.2±9.5</td>
</tr>
</tbody>
</table>

Table 4. Comparison with DAC [26] averaged over 5 trials (mean±std), where each trial uses different randomly sampled labeled data.

<table>
<thead>
<tr>
<th>Data</th>
<th>gate confidence</th>
<th>MAE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>label only</td>
<td>✓</td>
<td>145.39</td>
<td>257.31</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>144.48</td>
<td>255.33</td>
</tr>
<tr>
<td>label+unlabel</td>
<td>✓</td>
<td>138.52</td>
<td>242.26</td>
</tr>
</tbody>
</table>

Table 5. Ablation study on 5% label data of UCF-QNRF.

<table>
<thead>
<tr>
<th>Data</th>
<th>loss for unlabeled data</th>
<th>MAE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 loss</td>
<td></td>
<td>137.17</td>
<td>239.52</td>
</tr>
<tr>
<td>L2 w/ confidence</td>
<td></td>
<td>135.88</td>
<td>233.19</td>
</tr>
<tr>
<td>GL</td>
<td></td>
<td>125.32</td>
<td>214.96</td>
</tr>
<tr>
<td>GL w/ confidence(C-GL)</td>
<td></td>
<td>120.13</td>
<td>208.87</td>
</tr>
</tbody>
</table>

Table 6. Ablation study on semi-supervised counting when using different density-map localization methods. (mean±std).

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label only</td>
<td>138.52±10.65</td>
<td>242.26±16.62</td>
</tr>
<tr>
<td>GMM [14]</td>
<td>126.67±7.41</td>
<td>217.00±16.17</td>
</tr>
<tr>
<td>OT-M (ours)</td>
<td>120.13±7.34</td>
<td>208.87±11.65</td>
</tr>
</tbody>
</table>

Figure 7. The localization performance on unlabeled data improves during training. (x-axis is the training epoch.)

sets the number of pseudo-points as the sum of the predicted density map, and then estimates their locations. Compared with GMM, OT-M obtains more accurate pseudo hard-label, which leads to counting models that can better capture information from the unlabeled, and thus improves the semi-supervised learning performance. Specifically, Fig. 7 shows the localization performance on the unlabeled data during semi-supervised training. Generally, the localization improves during training, while OT-M obtains the best localization accuracy, i.e., the most accurate hard pseudo-labels.

6. Conclusion

This paper presents a parameter-free crowd localization method on density map, the OT-M algorithm. OT-M alternates between two steps: in the OT-step, the transport plan between the current point map and the input density map is estimated; in the M-step, the point map is updated using the transport plan computed in the OT-step. The convergence of OT-M is analyzed both in theory and practice. Experiments also show that OT-M outperforms previous localization methods based on density maps, as well as recent point detection methods. Furthermore, we apply OT-M to semi-supervised counting to produce hard pseudo-labels, and we propose a confidence-weighted generalized loss for this task, which assigns lower confidence to unlabeled data with inconsistency between teacher’s labels and student’s predictions. Empirical results demonstrate that efficacy of our framework on several crowd counting datasets.

Acknowledgements. This work was supported by a Strategic Research Grant from City University of Hong Kong (Project No. 7005665).
References


