Learning to Count via Unbalanced Optimal Transport

Zhiheng Ma[1], Xing Wei[2], Xiaopeng Hong[3,4*], Hui Lin[3], Yunfeng Qiu[2], Yihong Gong[2]
[1]College of Artificial Intelligence, Xi'an Jiaotong University
[2]School of Software Engineering, Xi'an Jiaotong University
[3]School of Cyber Science and Engineering, Xi'an Jiaotong University
[4]Research Center for Artificial Intelligence, Peng Cheng Laboratory
mazhiheng@stu.xjtu.edu.cn, {weixing,hongxiaopeng}@mail.xjtu.edu.cn, {waitandwait,yfqiu2015}@stu.xjtu.edu.cn,
ygong@mail.xjtu.edu.cn

Abstract

Counting dense crowds through computer vision technology has attracted widespread attention. Most crowd counting datasets use point annotations. In this paper, we formulate crowd counting as a measure regression problem to minimize the distance between two measures with different supports and unequal total mass. Specifically, we adopt the unbalanced optimal transport distance, which remains stable under spatial perturbations, to quantify the discrepancy between predicted density maps and point annotations. An efficient optimization algorithm based on the regularized semi-dual formulation of UOT is introduced, which alternatively learns the optimal transportation and optimizes the density regressor. The quantitative and qualitative results illustrate that our method achieves state-of-the-art counting and localization performance.

1 Introduction

Crowd counting has particularly important practical application value in public security. It can also be applied to count vehicles (Onoro Rubio and López-Sastre 2016), cells (Lempitsky and Zisserman 2010), and animals (Marsden et al. 2018). It is challenging to count targets in complex scenarios with thousands of targets, severe occlusions, and huge scale variations. As a result, in modern crowd counting datasets, it is more feasible and labor-saving to provide coarse annotations, *e.g.*, points, than detailed annotations such as bounding boxes or even contours.

Most previous methods adopt detection (Liu et al. 2019b, 2018a) or density regression (Zhang et al. 2016; Sindagi and Patel 2017) to predict the total count. They encounter an obstacle that the required ground truth is unavailable, such as the bounding boxes for detection and the pixel-wise density map for density regression. To ease this problem, some detection-based methods introduce additional bounding-box annotations (Liu et al. 2018a) or generate pseudo bounding boxes (Liu et al. 2019b; Sam et al. 2020), and most density-regression based methods generate pseudo density maps by smoothing the sparse annotated points with Gaussian kernels. Nevertheless, previous literature has shown that the 'quality' of the generated supervisory signals greatly affects the counting performance (Zhang et al. 2016; Ma et al. 2019; Liu et al. 2019b). For example, the performance of density-regression-based methods is highly influenced by the Gaussian kernel size (Wan and Chan 2019; Zhang et al. 2016). Moreover, generating *pseudo* ground truth also has side effects. For example, an oversized Gaussian kernel will reduce the location precision of points and result in poor quality density maps (Idrees et al. 2018).

In this paper, for the first time, we consider the counting problem from the perspective of *measure theory*. Specifically, we formulate the collection of scattered annotated points by a discrete measure $v = \sum_{j=1}^{M} v_j \delta_{y_j}$, where $v_j = 1$, y_j is the 2D position of the point, and $\mathcal{Y} = \{y_1, y_2, \dots, y_M\}$ is the support set of this discrete measure. Similarly, a density map can also be formulated by a discrete measure supported on the pixel space $\mathcal{X}$. It is worth noticing that since only a small portion of pixels are annotated with points, the supports of these two measures, $\mathcal{X}$ and $\mathcal{Y}$, only sparsely overlap. Therefore, an essential question about the measure-theory-based optimization method is how to quantify differences between the measure for scattered points and the measure for dense density maps.

The key to answering this question is to define the distance between two measures with different supports. As it is ambiguous to annotate a person occupying an area with a single point, the desired distance should be robust to the ambiguity in annotations and remain stable under spatial perturbations of point annotations. This requirement is described mathematically as *weak convergence* (Genevay 2019; Sriperumbudur et al. 2012). According to this understanding, ℓ_2 distance, widely used in density-regression methods, is proved not to metrize weak convergence and is sensitive to spatial perturbations (further discussed in Sec. 3.2). As a result, it can only be applied to discrete measures with the same supports. One alternative is to use optimal transport (OT) distance, which has been proved to capture the underlying geometry of measures with different supports and metrize weak convergence (Santambrogio 2015). Nonetheless, The prerequisite of using OT is that the total mass of two input measures should be equal. Thus OT can not handle mass variations, which are common in practice.

To address these limitations, we propose to use unbalanced optimal transport (UOT) distance to quantify discrepancy between two measures. Compared to OT, UOT 're-

[*]Corresponding author

laxes' the hard marginal constraint to the marginal divergence penalty (Liero, Mielke, and Savaré 2018). More importantly, UOT metrizes the weak convergence and allows mass change during transportation. On this basis, we formulate crowd counting as a measure regression problem.

Specifically, we regard the scattered point ground truth and the predicted density map as two measures with different supports, and introduce the UOT distance to quantify their differences. A two-step algorithm based on the semi-dual regularized formulation of UOT is proposed for efficient optimization, which alternatively learns optimal transportation and minimizes UOT distance. Moreover, we investigate the effects of different cost functions and marginal divergence penalties for crowd counting. Extensive experiments show that our methods achieve state-of-the-art counting performance. Furthermore, our method outputs a much sharper density map, closer to the ground truth, and thus achieves superior localization performance than previous methods.

The contributions of this paper are summarized as follows:

- We formulate crowd counting as a *measure regression* problem for the first time and propose a novel method to minimize the distance between two measures with different supports.
- We propose an objective function to minimize the unbalanced optimal transport distance between scattered ground-truth points and predicted density maps.
- We design an efficient optimization algorithm of counting models based on the UOT's semi-dual regularized formulation.
- Our method achieves state-of-the-art counting performance and superior localization performance with sharp density predictions.

2 Related Work

2.1 Crowd Counting

Due to the great progress of the convolutional neural networks (CNN), most state-of-the-art counting methods are CNN based. The majority of them adopt the fully convolutional network (FCN) (Long, Shelhamer, and Darrell 2015), which predicts density maps and maintain spatial correspondences with input images. Diverse network structures are developed to cope with large scale variations of crowds. Mechanisms such as multi-scale feature fusions (Idrees et al. 2013; Zhang et al. 2016; Sindagi and Patel 2017; Liu et al. 2018c; Liu, S., and F. 2019; Ma et al. 2020), attention (Jiang et al. 2020; Miao et al. 2020; Zhang et al. 2019), negative correlation learning (Shi et al. 2018; Zhang et al. 2019), and knowledge distilling (He et al. 2021; Liu et al. 2020a) are widely used. There are several works aiming to improve the quality of the generated pseudo ground truth. For example, (Zhang et al. 2016) adopts scale-adaptive Gaussian kernels. (Wan and Chan 2019) adopts CNN to learn combination of Gaussian kernels. (Liu et al. 2019a) estimates the kernel sizes with the aid of super-pixel segmentation.

The distance to measure the discrepancy between the predicted density maps and the ground truth is rarely studied, most methods use ℓ_2 distance. Some of them are simultaneously supervised by surrogate tasks, such as segmentation (Zhao et al. 2019; Gao, Wang, and Li 2019; Shi, Mettes, and Snoek 2019; Liu et al. 2020b), depth estimation (Zhao et al. 2019; Lian et al. 2019) and perspective estimation (Liu, S., and F. 2019; Shi et al. 2019; Yan et al. 2019; Yang et al. 2020). Bayesian loss (Ma et al. 2019) builds a probabilistic model and takes the expected count of each person as supervision, (Laradji et al. 2018) counts objects by segmentation, and (Liu et al. 2019) counts objects by detection. The concurrent work (Wang et al. 2020a) trains density regressor by matching normalized distributions through optimal transport.

2.2 (Un)balanced Optimal Transport

Optimal transport (OT) distances have recently been widely used in machine learning and deep learning, such as generative models (Arjovsky, Chintala, and Bottou 2017; Genevay, Peyré, and Cuturi 2018), image retrieval (Rubner, Tomasi, and Guibas 2000; Pele and Werman 2009), and domain adaption (Courty et al. 2017; Shen et al. 2018), to name a few. Calculating the exact OT is costly and suffers from the curse of dimensionality (Lee and Sidford 2014; Mérigot 2011). To deal with the scalability issue, (Cuturi 2013) introduces an entropic regularized OT and solves it by the Sinkhorn algorithm (Sinkhorn 1967). Reviews of OT can be found in (Kolouri et al. 2016; Peyré, Cuturi et al. 2019; Genevay 2019).

In comparison to OT, there are relatively fewer studies on the application of unbalanced optimal transportation (UOT). (Frogner et al. 2015) adopts UOT to train multi-label classification models. (Wang et al. 2020c) uses UOT to measure distances between documents of different lengths. (Chizat et al. 2018b) proposes a UOT-based fast approximate color transfer algorithm. (Yang and Uhler 2019) proposes a training algorithm of GAN. Several different formulations are proposed to extend the theory of OT to measures of unbalanced mass (Pham et al. 2020; Chizat et al. 2018c,a; Kondratyev et al. 2016; Liero, Mielke, and Savaré 2018; Bourne, Schmitzer, and Wirth 2018; Séjourné et al. 2019). A review of UOT can be found in (Chizat 2017).

3 The Proposed Method

In this section, we first elaborate on the distinction between the previous density-to-density training algorithm based on ℓ_2 distance and the density-to-point measure regression based on unbalanced optimal transport (UOT). Then, we review the basics of (un)balanced optimal transport. Finally, we introduce the optimization algorithm based on the semi-dual regularized formulation of UOT.

3.1 Problem Setup and Notation

Let $u = \sum_{i=1}^{N} u_i \delta_{x_i}, u_i = U_\theta(x_i; I)$ represent the predicted density map (density measure), where U_θ is the density regressor with the trainable parameter θ. The density regressor can be any fully convolutional networks (FCNs), which maintain the spatial correspondence between inputs and outputs. x_i is the 2D spatial position of the i_{th} pixel. $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ is the support of the density measure, where N is the total pixel number.

Let $v = \sum_{j=1}^{M} v_j \delta_{y_j}$ be the annotated points (point measure), where v_j is the mass carried with the j_{th} point, which equals to one in crowd counting ($v_j = 1$). y_j is the corresponding 2D spatial position. $\mathcal{Y} = \{y_1, y_2, \ldots, y_M\}$ is the support of the point measure, where M is the total point number. Since only a small portion of pixels is labeled with points, $\mathcal{X}$ and $\mathcal{Y}$ only sparsely overlap, and we have $M < N$.

3.2 Motivation

ℓ_2 distance is used in most of previous density-regression-based methods, which is easy to calculate but has two major drawbacks. **First**, ℓ_2 distance is unstable to spatial perturbation of annotated points, because it can not metrize *weak convergence* (Genevay 2019; Sriperumbudur et al. 2012). For example, $\ell_2(\delta_0, \delta_{1/n})$ is the distance between two Dirac masses supported on 0 and $1/n$, respectively. When n approaches infinity, a distance metrizing weak convergence should equal to zero. However, $\ell_2(\delta_0, \delta_{1/n})$ is still equal to a non-zero constant (*i.e.*, $\sqrt{2}$). Thus, even a small disturbance of annotated points could cause considerable changes in ℓ_2 distance. Specifically, weak convergence (Genevay 2019) is defined as follows: a measure α_n weakly converges to α ($\alpha_n \rightharpoonup \alpha$) if and only if $\int f(x)\mathrm{d}\alpha_n = \int f(x)\mathrm{d}\alpha$ for any continuous bounded function f; a distance D metrizes weak convergence if and only if $D(\alpha_n, \alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha$. **Second**, ℓ_2 distance can only be applied to discrete measures with the same supports. To alleviate these two shortcomings, previous methods transform the point measure v to the pseudo-ground-truth density measure $\bar{u}$ which shares the same support $\mathcal{X}$ with u:

$$\bar{u} = \sum_{i=1}^{N} \bar{u}_i \delta_{x_i}, \qquad \bar{u}_i = \sum_{j=1}^{M} \mathcal{N}(x_i; y_j, \sigma^2 \mathbf{1}_{2\times 2}), \tag{1}$$

where $\mathcal{N}(x_i; y_j, \sigma^2 \mathbf{1}_{2\times 2})$ is the Gaussian distribution whose mean is y_j and variance is σ. The density-to-density regression method is defined as follows:

$$\ell_2^2(u, \bar{u}) = \sum_{i=1}^{N} (u_i - \bar{u}_i)^2. \tag{2}$$

Although the use of pseudo-ground-truth density maps can stabilize ℓ_2 distance to some extent, it is just palliative and even brings new problems. First, there is no theoretical guide to determine the size of the Gaussian kernel. Second, Gaussian smoothing will reduce the location precision of points. In this paper, we try to solve this problem from the root cause, *i.e.*, directly minimize the distance between predicted density maps and ground-truth point annotations. The appropriate distance should allow input measures with different supports and unequal total mass as well as metrize weak convergence. Unbalanced optimal transport (UOT) distance is proved to meet all the requirements (Chizat 2017). In the following sections, we first review the basics of UOT, then introduce an efficient optimization algorithm based on its semi-dual regularized formulation.

3.3 Learning to Count via UOT

Preliminary. Optimal transport (OT) represents the minimum cost when pushing the mass in measure u to match that in measure v, $u \in \mathcal{M}_+(\mathcal{X})$ and $v \in \mathcal{M}_+(\mathcal{Y})$. $\mathcal{M}_+(\mathcal{X})$ represents the space of non-negative finite random measures on support $\mathcal{X}$, and $\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$ is a random non-negative measure on the product space. The Kantorovich formulation of OT is defined by:

$$\begin{aligned} &W(u, v) = \sup_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\mathrm{d}\pi(x, y), \\ &s.t.\ P_{\#}^{\mathcal{X}}\pi = u, P_{\#}^{\mathcal{Y}}\pi = v \end{aligned} \tag{3}$$

where $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ is the lower semi-continuous cost function. π is the joint measure having u and v as its marginals. $P^{\mathcal{X}}(x, y) = x$ is the projection map. $\#$ represents the push forward operator.

One major limitation of OT is the total mass of the input measures should be equal, otherwise, there is no feasible solution. It means that OT can not handle total mass variations. Therefore, we adopt unbalanced optimal transport (UOT) allows mass change during transportation:

$$\begin{aligned} W^{ub}(u, v) = &\sup_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\mathrm{d}\pi(x, y) \\ &+ \mathcal{F}_\varphi(P_{\#}^{\mathcal{X}}\pi | u) + \mathcal{F}_\varphi(P_{\#}^{\mathcal{Y}}\pi | v). \end{aligned} \tag{4}$$

As can be seen, the hard marginal constraints of OT are replaced by the "soft" marginal divergence penalty $\mathcal{F}_\varphi$:

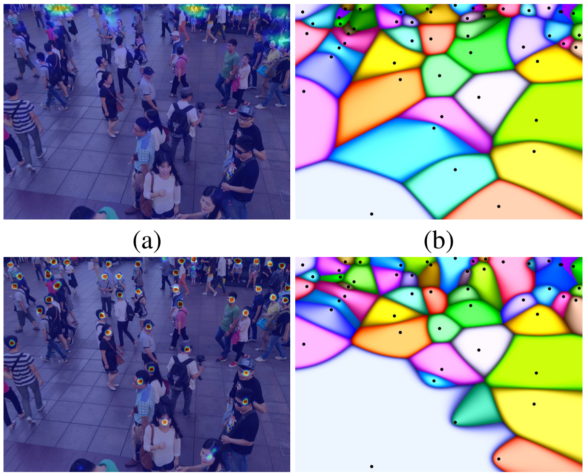
$$\mathcal{F}_\varphi(\rho|u) = \begin{cases} \int_{\mathcal{X}} \varphi(\frac{d\rho}{du}(x))du(x) & \text{if } \rho \ll u \\ \infty & \text{otherwise} \end{cases}, \tag{5}$$

where $\varphi : [0, \infty) \to [0, \infty)$ is a convex and lower semi-continuous function with $\lim\limits_{x\to\infty} \frac{\varphi(x)}{x} = \infty$ and $\varphi(1) = 0$. $\rho \ll u$ indicates the measure ρ is absolutely continuous *w.r.t.* the measure u. Obviously, classical OT is a special case of UOT when we set $\varphi(\rho|u) = 0$ if $\rho = u$ and ∞ for otherwise.

Semi-dual Regularized Optimization Algorithm. To calculate the UOT distance efficiently, we introduce an entropic regularizer (Cuturi 2013), which results in a smoothed transport map and a convex problem.

The primer formulation of regularized UOT on discrete measures is defined as follows:

$$\begin{aligned} W_\varepsilon^{ub}(u, v) &= \min_{\pi \in \mathbb{R}_+^{N\times M}} \mathcal{E}(\pi, u, v), \\ \mathcal{E}(\pi, u, v) &= \sum_{i=1}^{N}\sum_{j=1}^{M} c(x_i, y_j)\pi_{i,j} + \\ &\quad \mathcal{F}_\varphi(\pi \mathbb{1}_M | u) + \mathcal{F}_\varphi(\pi^\intercal \mathbb{1}_N | v) - \varepsilon\mathcal{H}(\pi), \\ \mathcal{H}(\pi) &= -\sum_{i=1}^{N}\sum_{j=1}^{M} (\pi_{i,j} \ln \pi_{i,j} - \pi_{i,j} + 1), \end{aligned} \tag{6}$$



(a) (b)

(c) (d)

Figure 1: Visualization of density maps and transport maps. (a) and (c) are the predicted density maps in the Epoch 1 and the Epoch 100, respectively, and (b) and (d) are their corresponding transport maps. For visualization of density maps, the warmer color represents the higher density value. For visualization of transport maps, annotated points are in black, and each person is assigned with a unique color. The color of each pixel is determined by its maximum transport target, and the brightness is determined by the corresponding value. As can be seen, the transport map is jointly determined by the predicted density map and the annotated points.

where $\mathbb{1}_N = (1, 1, \ldots, 1)^\intercal \in \mathbb{R}^N$. Most methods solve Eq. 6 by transforming it into a dual formulation and optimizing two variables ($\alpha \in \mathbb{R}^N$ and $\beta \in \mathbb{R}^M$) through Sinkhorn iterations (Sinkhorn 1967). With noticing that the annotated point number M is much smaller than the pixel number N, we further eliminate $\alpha \in \mathbb{R}^N$ through c-transform by Eq. 8, which results in the semi-dual formulation which only needs to optimize $\beta \in \mathbb{R}^M$:

$$\begin{aligned} W_\varepsilon^{ub}(u, v) &= \max_{\beta \in \mathbb{R}^M} \mathcal{G}(\beta, u, v), \\ \mathcal{G}(\beta, u, v) &= -\sum_{i=1}^{N} \varphi^*(-\alpha_i^\varepsilon) u_i - \sum_{j=1}^{M} \varphi^*(-\beta_j) v_j, \end{aligned} \tag{7}$$

$$\alpha_i^\varepsilon = \begin{cases} \min\limits_j c(x_i, y_j) - \beta_j & \varepsilon = 0 \\ -\varepsilon \ln \left(\sum\limits_{j=1}^{M} e^{\frac{\beta_j - c(x_i, y_j)}{\varepsilon}} \right) & \varepsilon > 0 \end{cases}, \tag{8}$$

where φ^* is the Frenchel-Legendre conjugate of the penalty function φ, which can be derived as follows:

$$\varphi^*(z) = \sup_{x >= 0} \left(z \cdot x - \varphi(x) \right). \tag{9}$$

In our experiments, we adopt the Kullback-Leibler (KL) divergence as the penalty function:

$$\varphi(x) = \begin{cases} x \ln(x) - x + 1 & x > 0 \\ 1 & x = 0 \end{cases} \qquad \varphi^*(z) = e^z - 1. \tag{10}$$

And we adopt the p-norm distance between the positions of annotated points and the positions of pixels of the density map as the cost function:

$$c(x_i, y_j) = \ell_p(x_i, y_j). \tag{11}$$

The semi-dual formulation is an unconstrained, differentiable optimization problem. Therefore, it can be solved by gradient-descent based algorithms. Specifically, we adopt L-BFGS quasi-Newton method (Liu and Nocedal 1989) in our experiments. The relationship between the primer and the semi-dual formulation is given by: $\hat{\pi}_{i,j} = \exp \frac{\hat{\alpha}_i^\varepsilon + \hat{\beta}_j - c(x_i, y_j)}{\varepsilon}$, where $\hat{\beta}$ and $\hat{\alpha}^\varepsilon$ are the optimal solutions of the semi-dual formulation. We visualize the predicted density map and the transport map in Fig. 1. As can be seen, the transport map is jointly determined by the predicted density map and the annotated points.

Finally, we propose a measure regression algorithm based on semi-dual regularized UOT (Eq. (7)), which alternatively optimizes the random variable β to find the optimal transport and the density regressor's parameter θ to minimize $W_\varepsilon^{ub}(u, v)$. Intuitively, mass of predicted density measures is adjusted and pushed to annotated points. Details of the algorithm are shown in Algorithm 1. When there are no point annotations in the image, we directly regress the total mass of the predicted density map to zero.

Algorithm 1: Unbalanced Optimal Transport From Density Predictions to Point Annotations

```
Input: current training step $t$, density estimator $U$
       with parameter $\theta$, input image $I$ with the
       ground-truth point measure $v$
Output: optimized density estimator $U$ with
        parameter $\hat{\theta}$
Initialize $\theta^{(1)}$;
for $t = 1, \ldots, T$ do
    $u^{(t)} = U_{\theta^{(t)}}(I^{(t)})$;
    Initialize $k = 1, \beta^{(t,k)} = \mathbb{O}_{M^t}$;
    // $\mathbb{O}_{M^t} \in \mathbb{R}^{M^t}$ is a zero vector
    repeat
        $\mathcal{G}^{(t,k)} = \mathcal{G}(\beta^{(t,k)}, u^{(t)}, v^{(t)})$;
        // Maximize $\mathcal{G}$ by optimizing $\beta$
        Update $\beta^{(t,k+1)}$ using L-BFGS;
        Update $k := k + 1$;
    until $\mathcal{G}^{(t,k)}$ has converged;
    $\hat{\beta}^{(t)} = \beta^{(t,k)}, W_\varepsilon^{ub} = \mathcal{G}(\hat{\beta}^{(t)}, u^{(t)}, v^{(t)})$;
    // Minimize $W_\varepsilon^{ub}$ by optimizing $\theta$
    Update $\theta^{(t+1)}$ using Adam;
end
return $\hat{\theta} = \theta^{(t+1)}$
```

4 Experimental Results

We conduct extensive experiments on the four largest crowd counting benchmarks to verify the effectiveness of the proposed method on both counting and localization tasks. We also study the impacts of different cost functions and marginal divergence penalties.

4.1 Implement Detail

We adopt the same network structure (VGG-19 truncated at the last pooling layer) used in Bayesian loss (BL) (Ma et al. 2019). For optimization, we set the learning rate of L-BFGS and Adam to 1.0 and 10^{-5}, respectively, and ε to 0.01. The principle to set ε is simple. A smaller ε leads to a closer approximation to the original UOT, but the convergence speed is slower. Therefore, ε should be as small as possible when the convergence rate is acceptable. We find that the convergence rate is too slow when $\varepsilon <= 0.001$, then we choose a moderate value, 0.01. Random crop and random horizontal flip are applied to augment input images. All the experiments are conducted on a single GPU card (Pascal Titan X), and our code is implemented with Pytorch.

4.2 Datasets

ShanghaiTech (Zhang et al. 2016), UCF-QNRF (Idrees et al. 2018), JHU-CROWD++ (Sindagi, Yasarla, and Patel 2020), NWPU-CROWD (Wang et al. 2020b), which are currently the largest and most diverse datasets, are used through out our experiments.

ShanghaiTech (Zhang et al. 2016) consists of Part A and Part B. Part A is composed of 482 images scraped from the Web, with 244,167 point annotations. The training set includes 300 images and the remaining 182 images are used for testing. Part B contains 716 fixed-resolution images taken from busy streets, with 88,498 point annotations. There are 400 images for training and the remaining 316 for testing.

UCF-QNRF (Idrees et al. 2018) contains 1,535 high-resolution images scraped from the Web, with 1.25 million point annotations. The training set includes 1,201 images and the remaining 334 images are used for testing.

JHU-CROWD++ (Sindagi, Yasarla, and Patel 2020) consists of 4,372 images with 1.51 million annotations. There are 2,272 images for training, 500 images for validation, and 1,600 images for testing. This dataset collects images under diverse scenarios and environmental conditions, such as different weathers and illumination.

NWPU-CROWD (Wang et al. 2020b) contains 5,109 images with 2.13 million annotations. There are 3,109 images for training, 500 images for validation, and 1,500 images for testing. This dataset introduces 351 distractors (scenes without people), which are similar to congested crowd scenes in terms of texture features.

4.3 Evaluation Metrics

Counting and localization are two critical tasks for crowd analysis. We evaluate both to compare different methods comprehensively.

Dataset	NWPU-CROWD		JHU-CROWD++		UCF-QNRF		ShanghaiTech_A		ShanghaiTech_B	
Method	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
ℓ_2	120.7	463.5	81.7	304.5	106.8	183.7	68.6	110.1	8.5	13.9
BL	105.4	454.2	75.0	299.1	88.7	154.8	62.8	101.8	7.7	12.7
UOT	**87.8**	**387.5**	**60.5**	**252.7**	**83.3**	**142.3**	**58.1**	**95.9**	**6.5**	**10.2**

Table 1: Comparisons with baseline methods sharing the same backbone (VGG-19).

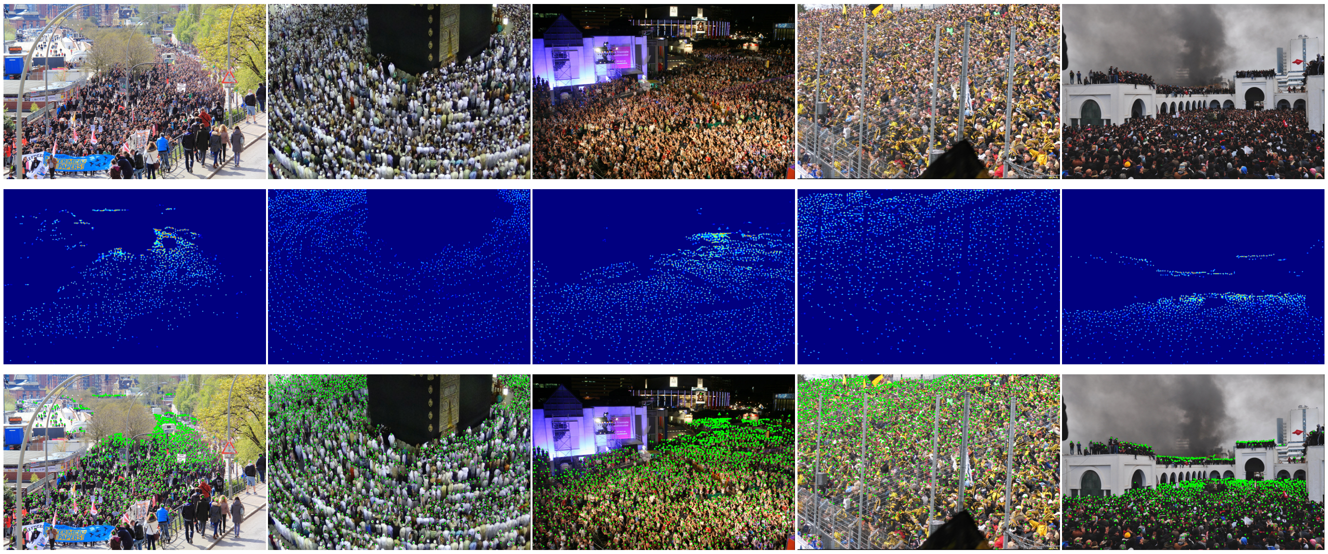
Counting performance is measured by three metrics (Zhang et al. 2016; Idrees et al. 2018; Wang et al. 2020b): Mean Absolute Error, $MAE = \frac{1}{K}\sum_{k=1}^{K}|N_k - C_k|$; Mean Squared Error, $MSE = (\frac{1}{K}\sum_{k=1}^{K}|N_k - C_k|^2)^{\frac{1}{2}}$; and Normalized Absolute Error, $NAE = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{N_k - C_k}{N_k}\right|$. K is the number of test images, N_k and C_k are the ground-truth count and the predicted count for the k-th image, respectively. The total count is obtained by summing the predicted density map $C_k = \sum_{i=1}^{N} U_{\hat{\theta}}(x_i, I_k)$.

Localization performance is measured by precision and recall at various distance thresholds (1,2,3,...,100 pixels), which is adopted by previous crowd localization methods (Idrees et al. 2018; Liu, Weng, and Mu 2019; Wang et al. 2020b). For density-based methods, post-processing is required to convert density maps into point predictions. In our experiments, we adopt the non-maximal suppression (NMS) to find local peaks of density maps as the predicted points. Then, each ground-truth point is associated with its nearest prediction through greedy one-to-one matching. If the distance between the matching pair is within the distance threshold, we predicate the matching is successful. The successfully matched predicted points are true positive (TP), the remaining predicted points are false positive (FP), and the unmatched annotated points are false negative (FN). Finally, The overall localization performance is measured by average precision (AP), average recall (AR), and F-measure.

4.4 Counting Evaluation

Comparison with baseline methods. We fairly compare three training algorithms with the same network structure (VGG-19), *i.e.*, ℓ_2 (density-to-density), BL (density-to-count), and UOT (density-to-point). The generation of pseudo-ground-truth density maps follows previous methods (Zhang et al. 2016; Wang et al. 2020b). Results of ℓ_2 and BL are reported from (Ma et al. 2019; Sindagi, Yasarla, and Patel 2020; Wang et al. 2020b). As shown in Tab. 1, UOT consistently outperforms the other two training methods on all benchmark datasets, which strongly proves the effectiveness of our method. Compared to the previous state-of-the-art training method BL, UOT reduces MAE and MSE by 17.6 and 66.7, respectively, on NWPU-CROWD.

Comparison with state-of-the-art methods. We extensively compare UOT with other state-of-the-art methods on four largest benchmark datasets. Their code has been officially released or reproduced by third parties. Experimental results are illustrated in Tab. 2-4, and the highlights can be summarized as follows: 1) UOT achieves the most advanced



(a) GT: 1557 Pre: 1503.2 (b) GT: 1798 Pre: 1885.8 (c) GT: 2414 Pre: 2437.5 (d) GT: 1408 Pre: 1785.7 (e) GT: 1454 Pre: 1395.1

Figure 2: Visualization of density maps and localization results. The first row: input images; The second row: predicted density maps; The third row: predicted locations of people (denoted by green points).

Category	Overall		Low		Median		High		Weather	
Method	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN (Zhang et al. 2016)	188.9	483.4	97.1	192.3	121.4	191.3	618.6	1166.7	330.6	852.1
CSRNet (Li, Zhang, and Chen 2018)	85.9	309.2	27.1	64.9	43.9	71.2	356.2	784.4	141.4	640.1
SANet (Cao et al. 2018)	91.1	320.4	17.3	37.9	46.8	69.1	397.9	817.7	154.2	685.7
CAN (Liu, S., and F. 2019)	100.1	314.0	37.6	78.8	56.4	86.2	384.2	789.0	155.4	617.0
SFCN (Wang et al. 2019)	77.5	297.6	16.5	55.7	38.1	59.8	341.8	758.8	122.8	606.3
DSSINet (Liu et al. 2019a)	133.5	416.5	53.6	112.8	70.3	108.6	525.5	1047.4	229.1	760.3
MBTTBF (Sindagi and Patel 2019)	81.8	299.1	19.2	58.8	41.6	66.0	352.2	760.4	138.7	631.6
LSC-CNN (Sam et al. 2020)	112.7	454.4	10.6	31.8	34.9	55.6	601.9	1,172.2	178.0	744.3
BL (Ma et al. 2019)	75.0	299.9	**9.5**	28.1	30.0	49.8	289.5	659.0	124.9	614.7
CG-DRCN-CC (Sindagi, Yasarla, and Patel 2020)	71.0	278.6	14.0	42.8	35.0	53.7	314.7	712.3	120.0	**580.8**
UOT	**60.5**	**252.7**	11.2	**26.2**	**28.7**	**45.3**	**274.1**	**648.2**	**114.9**	610.7

Table 2: Counting results on JHU-CROWD++. JHU-CROWD++ divides test set into following fine-grained subsets: (1) Low: images with 0 to 50 people, (2) Medium: images with 51 to 500 people, (3) High: images containing more than 500 people, and (4) Weather: degraded images.

performance on NWPU-CROWD, JHU-CROWD++, UCF-QNRF, and ShanghaiTech_B, and also achieves competitive performance on ShanghaiTech_A. 2) UOT performs well on all crowding levels. It improves NAE from 0.203 to 0.185 on NWPU-CROWD.

4.5 Localization Evaluation

Highly congested crowd images contain severe occlusions among individuals, and the separation between adjacent annotated points may even be only a few pixels. Therefore, the "sharpness" of the predicted density maps is crucial for the localization task. As visualized in Fig. 2, UOT predicts sharp density maps (the second row) which benefits distinguishing individuals from congested crowds. The predicted locations (the third row) are denoted with green dots. As shown, UOT is able to locate people with severe overlaps and is robust to changes in size, luminance, and appearance. The quantitative results of localization are reported in Tab. 5, showing that UOT outperforms other methods on UCF-QNRF. In particular, UOT improves the F-measure from 70.85% to 74.75%, compared to BL.

4.6 Discussion

In this section, we study the effects of different cost functions and penalty functions on crowd counting. In addition to p-norm distance that allows transport within the whole image, we also study the Wasserstein–Fisher–Rao distance,

Category	Overall			Sense Level (MAE)		Luminance (MAE)	
Method	MAE	MSE	NAE	Avg.	$S_0 \backsim S_4$	Avg.	$L_0 \backsim L_2$
MCNN (Zhang et al. 2016)	232.5	714.6	1.063	1171.9	356.0 / 72.1 / 103.5 / 509.5 / 4818.2	220.9	472.9 / 230.1 / 181.6
SANET (Cao et al. 2018)	190.6	491.4	0.991	716.3	432.0 / 65.0 / 104.2 / 385.1 / 2595.4	153.8	254.2 / 192.3 / 169.7
DECIDENET (Liu et al. 2018b)	264.9	759.0	1.770	1242.5	443.0 / 125.5 / 140.5 / 461.5 / 5036.6	313.6	464.2 / 267.4 / 209.1
CSRNET (Li, Zhang, and Chen 2018)	121.3	387.8	0.604	**522.7**	176.0 / 35.8 / 59.8 / 285.8 / **2055.8**	112.0	232.4 / 121.0 / 95.5
CAN (Liu, S., and F. 2019)	106.3	**386.5**	0.295	612.2	82.6 / 14.7 / 46.6 / 269.7 / 2647.0	102.1	222.1 / 104.9 / 82.3
SCAR (Gao, Wang, and Yuan 2019)	110.0	495.3	0.288	718.3	122.9 / 16.7 / 46.0 / 241.7 / 3164.3	102.3	223.7 / 112.7 / 73.9
BL (Ma et al. 2019)	105.4	454.2	0.203	750.5	66.5 / 8.7 / 41.2 / 249.9 / 3386.4	115.8	293.4 / 102.7 / 68.0
SFCN+ (Wang et al. 2020b)	105.7	424.1	0.254	712.7	**54.2** / 14.8 / 44.4 / 249.6 / 3200.5	106.8	245.9 / 103.4 / 78.8
UOT	**87.8**	387.5	**0.185**	566.5	80.7 / **7.9** / **36.3** / **212.0** / 2495.4	**95.2**	240.3 / **86.4** / **54.9**

Table 3: Counting results on NWPU-CROWD. NWPU-CROWD divides test set into following fine-grained subsets: (1) S_0: images without people, but containing textures similar to the crowd. (2) S_1: images with 1 to 100 people, (3) S_2: images with 101 to 500 people, (4) S_3: images with 501 to 5000 people, and (5) S_4: images containing more than 5000 people. There are three more subsets based on images' average luminance values in the YUV color space, which are, (6) L_0: luminance value between $[0, 0.25]$, (7) L_1: luminance value between $(0.25, 0.5]$, and (8) L_2: luminance value between $(0.5, 0.75]$.

Dataset	PartA		PartB		UCF-QNRF	
Method	MAE	MSE	MAE	MSE	MAE	MSE
MCNN (Zhang et al. 2016)	110.2	173.2	26.4	41.3	277	426
SWITCH-CNN (Babu Sam 2017)	90.4	135.0	21.6	33.4	228	445
CAN (Liu, S., and F. 2019)	62.3	100.0	7.8	12.2	107	183
SFCN (Wang et al. 2019)	64.8	107.5	7.6	13.0	102.0	171.4
DSSINET (Liu et al. 2019a)	60.6	96.1	6.9	10.4	99.1	159.2
BL (Ma et al. 2019)	62.8	101.8	7.7	12.7	88.7	154.8
S-DCNET (Xiong et al. 2019)	58.3	95.0	6.7	10.7	104.4	176.1
ASNET (Jiang et al. 2020)	**57.8**	**90.1**	-	-	91.6	159.7
UOT	58.1	95.9	**6.5**	**10.2**	**83.3**	**142.3**

Table 4: Counting results on ShanghaiTech and UCF-QNRF.

Method	AP	AR	F-measure
MCNN (Zhang et al. 2016)	59.95%	63.50%	61.66%
RESNET74 (He et al. 2016)	61.60%	66.90%	64.14%
DENSENET63 (Huang et al. 2017)	70.91%	58.10%	63.87%
CL-CNN (Idrees et al. 2018)	**75.80%**	59.75%	66.82%
BL (Ma et al. 2019)	66.85%	75.36%	70.85%
ℓ_2	74.75%	64.64%	69.33%
UOT	71.41%	**78.41%**	**74.75%**

Table 5: Localization results on UCF-QNRF.

$c(x_i, y_j)$	$\ell_1(x_i, y_j)$		$\ell_2(x_i, y_j)$		$WFR(x_i, y_j)$	
Penalty	MAE	MSE	MAE	MSE	MAE	MSE
KL	92.7	158.9	**83.3**	**142.3**	85.8	153.7
QR	91.6	158.3	87.1	150.4	89.5	152.1

Table 6: Effects of different cost functions and penalties on crowd counting (UCF-QNRF).

which prohibits long-distance transport:

$$WFR(x_i, y_j) = \begin{cases} -2\ln\cos(\frac{\ell_p(x_i, y_j)}{\delta}) & \ell_p(x_i, y_j) < \frac{\pi}{2}\delta \\ \infty & \text{otherwise} \end{cases}, \tag{12}$$

where transportation between masses with a distance greater than $\frac{\pi}{2}\delta$ is prohibited. For penalty function, we also study the Quadratic (QR) divergence besides KL divergence:

$$\varphi(x) = (x-1)^2, \qquad \varphi^*(z) = \begin{cases} \frac{z^2}{4} + z & z \geq -2 \\ -1 & z < -2 \end{cases}. \tag{13}$$

Results in Tab. 6 show that ℓ_2 distance with KL divergence achieves the best performance, and we use this combination in other experiments. It also proves that global transport is better than local transport on crowd counting.

5 Conclusions and Future Work

In this paper, we formulate crowd counting as a measure regression problem, which directly minimizes unbalanced optimal transport distance between predicted density maps and ground-truth point annotations. We also propose an efficient optimization algorithm based on UOT's semi-dual regularized formulation. This solution is straightforward and does not need to introduce any assumptions or other objective functions. Extensive experiments prove the effectiveness of this method on both crowd counting and localization. A promising direction for future work is to adapt this method to other applications with point supervision, such as landmark detection and pose estimation.

Acknowledgements. This work is funded by National Key Research and Development Project of China under Grant No. 2020AAA0105600 and 2019YFB1312000, National Natural Science Foundation of China under Grant No. 62006183 and 62076195, and by China Postdoctoral Science Foundation under Grant No. 2020M683489.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. *ICML* .

Babu Sam, D. 2017. Switching Convolutional Neural Network for Crowd Counting. In *CVPR*.

Bourne, D. P.; Schmitzer, B.; and Wirth, B. 2018. Semi-discrete unbalanced optimal transport and quantization. *arXiv preprint arXiv:1808.01962* .

Cao, X.; Wang, Z.; Zhao, Y.; and Su, F. 2018. Scale Aggregation Network for Accurate and Efficient Crowd Counting. In *ECCV*.

Chizat, L. 2017. *Unbalanced optimal transport: Models, numerical methods, applications*. Ph.D. thesis, PSL Research University.

Chizat, L.; Peyré, G.; Schmitzer, B.; and Vialard, F.-X. 2018a. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics* 18(1): 1–44.

Chizat, L.; Peyré, G.; Schmitzer, B.; and Vialard, F.-X. 2018b. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation* 87(314): 2563–2609.

Chizat, L.; Peyré, G.; Schmitzer, B.; and Vialard, F.-X. 2018c. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis* 274(11): 3090–3123.

Courty, N.; Flamary, R.; Habrard, A.; and Rakotomamonjy, A. 2017. Joint distribution optimal transportation for domain adaptation. In *NeurIPS*, 3730–3739.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2292–2300.

Frogner, C.; Zhang, C.; Mobahi, H.; Araya, M.; and Poggio, T. A. 2015. Learning with a Wasserstein loss. In *NeurIPS*, 2053–2061.

Gao, J.; Wang, Q.; and Li, X. 2019. PCC Net: Perspective Crowd Counting via Spatial Convolutional Network. *TCSVT* .

Gao, J.; Wang, Q.; and Yuan, Y. 2019. SCAR: Spatial-/channel-wise attention regression networks for crowd counting. *Neurocomputing* 363: 1–8.

Genevay, A. 2019. *Entropy-regularized optimal transport for machine learning*. Ph.D. thesis, Paris Sciences et Lettres.

Genevay, A.; Peyré, G.; and Cuturi, M. 2018. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 1608–1617.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, Y.; Ma, Z.; Wei, X.; Hong, X.; Ke, W.; and Gong, Y. 2021. Error-Aware Density Isomorphism Reconstruction for Unsupervised Cross-Domain Crowd Counting. In *AAAI*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *CVPR*, 2261–2269.

Idrees, H.; Saleemi, I.; Seibert, C.; and Shah, M. 2013. Multi-source Multi-scale Counting in Extremely Dense Crowd Images. In *CVPR*, 2547–2554.

Idrees, H.; Tayyab, M.; Athrey, K.; Zhang, D.; Al-Maadeed, S.; Rajpoot, N.; and Shah, M. 2018. Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds. In *ECCV*.

Jiang, X.; Zhang, L.; Xu, M.; Zhang, T.; Lv, P.; Zhou, B.; Yang, X.; and Pang, Y. 2020. Attention Scaling for Crowd Counting. In *CVPR*, 4706–4715.

Kolouri, S.; Park, S.; Thorpe, M.; Slepčev, D.; and Rohde, G. K. 2016. Transport-based analysis, modeling, and learning from signal and data distributions. *arXiv preprint arXiv:1609.04767* .

Kondratyev, S.; Monsaingeon, L.; Vorotnikov, D.; et al. 2016. A new optimal transport distance on the space of finite Radon measures. *Advances in Differential Equations* 21(11/12): 1117–1164.

Laradji, I. H.; Rostamzadeh, N.; Pinheiro, P. H. O.; Vázquez, D.; and Schmidt, M. 2018. Where are the Blobs: Counting by Localization with Point Supervision. In *ECCV*.

Lee, Y. T.; and Sidford, A. 2014. Path finding methods for linear programming: Solving linear programs in o (vrank) iterations and faster algorithms for maximum flow. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, 424–433. IEEE.

Lempitsky, V.; and Zisserman, A. 2010. Learning to count objects in images. In *NeurIPS*.

Li, Y.; Zhang, X.; and Chen, D. 2018. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In *CVPR*.

Lian, D.; Li, J.; Zheng, J.; Luo, W.; and Gao, S. 2019. Density Map Regression Guided Detection Network for RGB-D Crowd Counting and Localization. In *CVPR*.

Liero, M.; Mielke, A.; and Savaré, G. 2018. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae* 211(3): 969–1117.

Liu, C.; Weng, X.; and Mu, Y. 2019. Recurrent Attentive Zooming for Joint Crowd Counting and Precise Localization. In *CVPR*.

Liu, D. C.; and Nocedal, J. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical programming* 45(1-3): 503–528.

Liu, J.; Gao, C.; Meng, D.; and Hauptmann, A. G. 2018a. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *CVPR*.

Liu, J.; Gao, C.; Meng, D.; and Hauptmann, A. G. 2018b. DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation. In *CVPR*.

Liu, L.; Chen, J.; Wu, H.; Chen, T.; Li, G.; and Lin, L. 2020a. Efficient Crowd Counting via Structured Knowledge Transfer. In *ACM MM*.

Liu, L.; Qiu, Z.; Li, G.; Liu, S.; Ouyang, W.; and Lin, L. 2019a. Crowd counting with deep structured scale integration network. In *ICCV*, 1774–1783.

Liu, L.; Wang, H.; Li, G.; Ouyang, W.; and Lin, L. 2018c. Crowd counting using deep recurrent spatial-aware network. In *IJCAI*.

Liu, W.; S., M.; and F., P. 2019. Context-Aware Crowd Counting. In *CVPR*.

Liu, Y.; Liu, L.; Wang, P.; Zhang, P.; and Lei, Y. 2020b. Semi-Supervised Crowd Counting via Self-Training on Surrogate Tasks. *ECCV* .

Liu, Y.; Shi, M.; Zhao, Q.; and Wang, X. 2019b. Point in, box out: Beyond counting persons in crowds. In *CVPR*.

Liu, Y.; Shi, M.; Zhao, Q.; and Wang, X. 2019. Point in, Box Out: Beyond Counting Persons in Crowds. In *CVPR*, 6462–6471.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, 3431–3440.

Ma, Z.; Wei, X.; Hong, X.; and Gong, Y. 2019. Bayesian Loss for Crowd Count Estimation With Point Supervision. In *ICCV*.

Ma, Z.; Wei, X.; Hong, X.; and Gong, Y. 2020. Learning Scales from Points: A Scale-Aware Probabilistic Model for Crowd Counting. In *ACM MM*, 220–228.

Marsden, M.; McGuinness, K.; Little, S.; Keogh, C. E.; and O'Connor, N. E. 2018. People, penguins and petri dishes: Adapting object counting models to new visual domains and object types without forgetting. In *CVPR*, 8070–8079.

Mérigot, Q. 2011. A Multiscale Approach to Optimal Transport. *Comput. Graph. Forum* 30(5): 1583–1592.

Miao, Y.; Lin, Z.; Ding, G.; and Han, J. 2020. Shallow Feature Based Dense Attention Network for Crowd Counting. In *AAAI*, 11765–11772.

Onoro Rubio, D.; and López-Sastre, R. J. 2016. Towards perspective-free object counting with deep learning. In *ECCV*.

Pele, O.; and Werman, M. 2009. Fast and robust earth mover's distances. In *ICCV*, 460–467. IEEE.

Peyré, G.; Cuturi, M.; et al. 2019. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning* 11(5-6): 355–607.

Pham, K.; Le, K.; Ho, N.; Pham, T.; and Bui, H. 2020. On Unbalanced Optimal Transport: An Analysis of Sinkhorn Algorithm. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 7673–7682. Virtual.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *IJCV* 40(2): 99–121.

Sam, D. B.; Peri, S. V.; Kamath, A.; Babu, R. V.; et al. 2020. Locate, Size and Count: Accurately Resolving People in Dense Crowds via Detection. *TPAMI* .

Santambrogio, F. 2015. Optimal transport for applied mathematicians. *Birkäuser, NY* 55(58-63): 94.

Séjourné, T.; Feydy, J.; Vialard, F.-X.; Trouvé, A.; and Peyré, G. 2019. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958* .

Shen, J.; Qu, Y.; Zhang, W.; and Yu, Y. 2018. Wasserstein distance guided representation learning for domain adaptation. *AAAI* .

Shi, M.; Yang, Z.; Xu, C.; and Chen, Q. 2019. Revisiting perspective information for efficient crowd counting. In *CVPR*.

Shi, Z.; Mettes, P.; and Snoek, C. G. 2019. Counting with focus for free. In *ICCV*, 4200–4209.

Shi, Z.; Zhang, L.; Liu, Y.; Cao, X.; Ye, Y.; Cheng, M.; and Zheng, G. 2018. Crowd Counting with Deep Negative Correlation Learning. In *CVPR*, 5382–5390.

Sindagi, V. A.; and Patel, V. M. 2017. Generating High-Quality Crowd Density Maps Using Contextual Pyramid CNNs. In *ICCV*.

Sindagi, V. A.; and Patel, V. M. 2019. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *ICCV*, 1002–1012.

Sindagi, V. A.; Yasarla, R.; and Patel, V. M. 2020. JHU-CROWD++: Large-Scale Crowd Counting Dataset and A Benchmark Method. *TPAMI* .

Sinkhorn, R. 1967. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly* 74(4): 402–405.

Sriperumbudur, B. K.; Fukumizu, K.; Gretton, A.; Schölkopf, B.; Lanckriet, G. R.; et al. 2012. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics* 6: 1550–1599.

Wan, J.; and Chan, A. 2019. Adaptive density map generation for crowd counting. In *ICCV*, 1130–1139.

Wang, B.; Liu, H.; Samaras, D.; and Nguyen, M. H. 2020a. Distribution Matching for Crowd Counting. *NeurIPS* .

Wang, Q.; Gao, J.; Lin, W.; and Li, X. 2020b. NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting. *TPAMI* .

Wang, Q.; Gao, J.; Lin, W.; and Yuan, Y. 2019. Learning from synthetic data for crowd counting in the wild. In *CVPR*, 8198–8207.

Wang, Z.; Zhou, D.; Yang, M.; Zhang, Y.; Rao, C.; and Wu, H. 2020c. Robust Document Distance with Wasserstein-Fisher-Rao metric. In *Proceedings of The 12th Asian Conference on Machine Learning*, 721–736.

Xiong, H.; Lu, H.; Liu, C.; Liu, L.; Cao, Z.; and Shen, C. 2019. From open set to closed set: Counting objects by spatial divide-and-conquer. In *ICCV*, 8362–8371.

Yan, Z.; Yuan, Y.; Zuo, W.; Tan, X.; Wang, Y.; Wen, S.; and Ding, E. 2019. Perspective-guided convolution networks for crowd counting. In *ICCV*, 952–961.

Yang, K. D.; and Uhler, C. 2019. Scalable Unbalanced Optimal Transport using Generative Adversarial Networks. In *ICLR*. OpenReview.net.

Yang, Y.; Li, G.; Wu, Z.; Su, L.; Huang, Q.; and Sebe, N. 2020. Reverse Perspective Network for Perspective-Aware Object Counting. In *CVPR*, 4374–4383.

Zhang, A.; Shen, J.; Xiao, Z.; Zhu, F.; Zhen, X.; Cao, X.; and Shao, L. 2019. Relational attention network for crowd counting. In *ICCV*, 6788–6797.

Zhang, L.; Shi, Z.; Cheng, M. M.; Liu, Y.; Bian, J. W.; Zhou, J. T.; Zheng, G.; and Zeng, Z. 2019. Nonlinear Regression via Deep Negative Correlation Learning. *TPAMI* .

Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; and Ma, Y. 2016. Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. In *CVPR*.

Zhao, M.; Zhang, J.; Zhang, C.; and Zhang, W. 2019. Leveraging Heterogeneous Auxiliary Tasks to Assist Crowd Counting. In *CVPR*.