

Direct Measure Matching for Crowd Counting

Hui Lin¹, Xiaopeng Hong^{1,4*}, Zhiheng Ma², Xing Wei³, Yunfeng Qiu³, Yaowei Wang⁴, Yihong Gong³

¹School of Cyber Science and Engineering, Xi’an Jiaotong University;

²College of Artificial Intelligence, Xi’an Jiaotong University;

³School of Software Engineering, Xi’an Jiaotong University;

⁴Pengcheng Laboratory, Shenzhen

waitandwait@stu.xjtu.edu.cn; hongxiaopeng@ieee.org; mazhiheng@stu.xjtu.edu.cn; weixing@mail.xjtu.edu.cn; yfqi2015@stu.xjtu.edu.cn; wangyw@pcl.ac.cn; ygong@mail.xjtu.edu.cn

Abstract

Traditional crowd counting approaches usually use Gaussian assumption to generate pseudo density ground truth, which suffers from problems like inaccurate estimation of the Gaussian kernel sizes. In this paper, we propose a new measure-based counting approach to regress the predicted density maps to the scattered point-annotated ground truth directly. First, crowd counting is formulated as a measure matching problem. Second, we derive a semi-balanced form of Sinkhorn divergence, based on which a Sinkhorn counting loss is designed for measure matching. Third, we propose a self-supervised mechanism by devising a Sinkhorn scale consistency loss to resist scale changes. Finally, an efficient optimization method is provided to minimize the overall loss function. Extensive experiments on four challenging crowd counting datasets namely ShanghaiTech, UCF-QNRF, JHU++ and NWPU have validated the proposed method.

1 Introduction

Crowd counting has become increasingly important in the fields of artificial intelligence and computer vision. It has been widely used in congestion estimation and crowd management. With the use of Convolutional Neural Networks (CNNs), crowd counting has achieved considerable success in recent years. However, due to complex images and coarse (point) annotations, the task is still challenging.

Existing crowd counting methods can be briefly categorized into two types. Detection based methods count the number of people by exhaustively detecting every individual in images [Liu *et al.*, 2019b] [Liu *et al.*, 2018]. Their applications are limited as they usually require additional annotations such as bounding boxes. Regression based methods regress the output to a *pseudo* density map by smoothing scattered annotated points with a fixed-size Gaussian kernel [Zhang *et al.*, 2016] [Sindagi and Patel, 2017] [Zhang *et al.*, 2018]. However, the quality of the generated pseudo

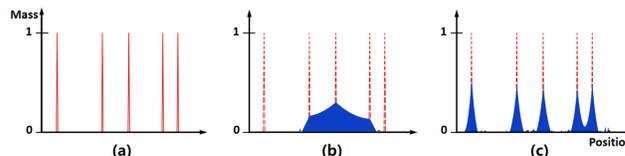


Figure 1: Comparisons of using Wasserstein distance (WD) and our proposed semi-balanced Sinkhorn divergence (SSD) as regressive loss. (a): The regression target. (b): The WD based regression results may shrink to a mass and suffers from *entropic bias*. (c): The SSD based regression output is sharp, well separated, and clearly centered at the annotated points.

maps highly depends on the settings of the Gaussian kernel size. It is shown that inappropriate sizes of Gaussian kernels will greatly impair the quality of density maps [Idrees *et al.*, 2018].

To solve these problems, in this paper, we regard crowd counting as a measure matching problem, based on the understanding that the scattered ground truth and the predicted density map can be expressed by a discrete point measure and a continuous measure, respectively.

It is essential to choose appropriate distance metric when matching two measures of different types. This may at first glance appear trivial as regularized Wasserstein distance is recently popular to calculate the discrepancy between measures. The regularized Wasserstein distance usually appends an entropic function as a regularization term to the overall objective function to relax the hard constraints during the assignment of optimal transport amount and reduce computational costs [Cuturi, 2013]. However, there exist two serious problems. First, the entropic term breaks the axiom of *Identity of indiscernibles*, i.e., $d(x, x) = 0$, which is the fundamental property of a metric space. Second, the quality of the optimization is severely sensitive to the parameter settings of the entropic term [Feydy *et al.*, 2019]. Inappropriate entropic parameters will cause the trainable measure shrink to a mass at the barycenter of the target. This phenomenon is often referred to as *entropic bias*, which is visualized in Figure 1.

Sinkhorn divergence can be an opinion to fix *the identity of indiscernibles* and the *hard constraints issues*. Sinkhorn divergence is with a self-correcting term. Thus it is more stable

*Corresponding author

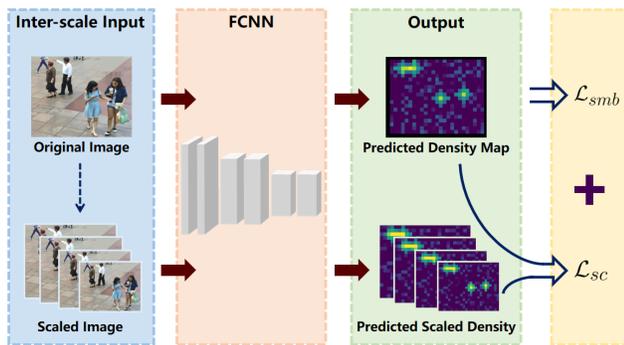


Figure 2: The framework of our proposed method S3. It has two compositions. First, the counting loss \mathcal{L}_{smb} directly measures the divergence from ground truth by the semi-balanced Sinkhorn. Second, the scale consistency loss \mathcal{L}_{sc} punishes the deviations in the measure space of various crowd scales.

under different entropic parameters and with a better interpolation meaning, compared to Wasserstein distance [Ramdas *et al.*, 2017]. Nevertheless, it is still constrained by the *equivalence of measures’ quantity*, which requires that the total amount of predicted density measure equals to the amount of ground truth measure. As a result, it cannot be directly applied to practical scenarios where the total mass of the predicted density map is different to those of the scattered annotation maps.

In this paper, we propose a new measure-theory based counting approach which directly regresses to point annotations, termed by semi-balanced Sinkhorn with scale consistency (S3). Firstly, to break the aforementioned limitations, we derive a semi-balanced form of the Sinkhorn distance and design a *semi-balanced Sinkhorn counting loss*. This new formulation relaxes the amount constraint and is fully in line with our problem assumption. Secondly, although the information about crowd scales has shown to be significant to counting [Zhang *et al.*, 2016] [Zeng *et al.*, 2017] [Cao *et al.*, 2018], it is lost when only the point annotations are available. To overcome this deficiency, we propose a self-supervised scale enhanced mechanism by using the inter-scale consistency in Sinkhorn distance and devise a *Sinkhorn scale consistency loss*. Thirdly, we derive the first-order conditions of the overall loss function and introduce the scaling iterations for efficient optimization. The framework of our method is depicted in Figure 2.

To evaluate our method, we have conducted extensive experiments and achieved promising results. The superior performance of our method is also demonstrated from the perspective of visualizations, where the outputs of the proposed method appear sharp and locate closed to annotated points.

The contributions of this paper are summarized as follows:

- We propose a new measure-matching based crowd counting approach, which can directly regress the dense density map to the scattered point annotations, without using Gaussian assumptions to generate poor-quality *pseudo* ground truth.
- We derive a semi-balanced form of Sinkhorn divergence

for computing the distance between two heterogeneous measures with different masses.

- We propose a self-supervised scale enhanced mechanism to improve the robustness against scale variations.
- Extensive experiments illustrate that our method achieves highly competitive counting performance.

2 Related Works

2.1 Crowd Counting

Crowd Counting has experienced rapid development due to the supports of a myriad of methods. Early papers tended to adopt detection of heads or bodies of crowd [Wu and Nevatia, 2005] but are limited by the high-density crowd congestion. Some works have introduced direct regressions with low-level features [Chen *et al.*, 2012] [Brostow and Cipolla, 2006]. More recently, deep CNN based crowd counting methods have achieved outstanding performances. [Zhang *et al.*, 2016] presents a multi-column CNN which regresses to the pseudo density map with adaptive Gaussian kernels. [Li *et al.*, 2018] applies a dilated network suitable for highly congested scenes. [Zhang *et al.*, 2019] proposes a relational attention network with exploration of interdependence of pixels. Moreover, methods based on segmentation [Sajid *et al.*, 2016], perspective estimation [Yan *et al.*, 2019], error estimation [He *et al.*, 2021] and multi-scale mechanisms [Zeng *et al.*, 2017] [Sindagi and Patel, 2017] [Liu *et al.*, 2019a] are proposed by latest papers, in order to break the limitations of perspective and scale variations of crowds. Furthermore, the method [Ma *et al.*, 2019] adopts Bayesian assumption and calculates with the expected count of pixels. Our method is distinct to most existing approaches. We treat crowd counting as a *measure matching* problem while others usually regard them as *density-map-matching* ones. Moreover, we introduce the measure theory based distance to gauge the distances of measures, which are beyond the scope of most existing crowd counting studies.

Recently, [Wang *et al.*, 2020a] tries to use optimal transport (OT) to measure the similarity between the normalized predicted and ground truth density map. But it is still constrained by the amount of density map and some unpleasing properties of traditional optimal transport.

2.2 Wasserstein and Sinkhorn Divergences

Wasserstein distance, *a.k.a.* Earth Mover’s Distance, provides an efficient way of calculating the distance between measures. It has many favorable properties, *e.g.*, convexity, tightness and the existence of optimal couplings [Villani, 2008], and thus has been widely used in many artificial intelligence applications these years. Some generative adversarial networks employ Wasserstein distance to generate computing functions with better properties [Arjovsky *et al.*, 2017]. It has also been extensively adopted in domain adaptation [Shen *et al.*, 2018a] and metric learning [Xu *et al.*, 2018].

Alternatively, Sinkhorn divergence eliminates the entropic bias and gathers the respective strengths between Wasserstein distance and MMD [Ramdas *et al.*, 2017] [Feydy *et al.*, 2019]. It was early proposed to solve the generative models [Genevay *et al.*, 2018]. [Séjourné *et al.*, 2019] extends it

to unbalanced optimal transport and elaborates on its attractive properties. To the best of our knowledge, there are no existing studies on using Sinkhorn distance for calculating between the distributions of two crowds.

3 The Proposed Method

In this section, we will detail our measure theory based semi-balanced Sinkhorn divergence with scale consistency. By first defining the problem as measure matching, we then explain the shortages of using traditional Wasserstein and Sinkhorn for evaluating the divergence. We elaborate the proposed semi-balanced Sinkhorn distance and use it as our counting loss. Then, we adopt inter-scale consistency mechanism and measure the deviations as scale loss. We combine these two losses for jointly regression. Finally, we will also investigate on computing solutions.

3.1 Problem Definition

Traditional method, which is based on pseudo density map, regards the counting problem as density regression. It adopts L_2 loss by generating the Gaussian-kernel density map with same pixel number as the output.

Let \mathcal{X}, \mathcal{Y} denote the 2-D supports of estimated density map and ground truth respectively. By redefining the problem as measure matching, from the perspective of measure theory, we represent the ground truth by $\beta = \sum_{j=1}^M \beta(y_j) \delta_{y_j}$. Since j represents an annotated point (a person in crowd counting), $\beta(y_j) = 1$, $m(\beta) = \sum_{j=1}^M \beta(y_j)$ equals to the number of people M . $y_j \in \mathcal{Y}$ signals the location of person j , and δ_{y_j} is a unit Dirac located at y_j .

Similar to the point measure, the output of density regressor can be defined as a continuous measure:

$$\alpha = \int_{\mathcal{X}} \alpha(x) dx, \text{ where } \alpha(x) = R_{\theta}(x; I).$$

R_{θ} is the density regressor with the trainable parameter θ and $\alpha(x)$ is nonnegative.

Consequently, the objective of regression is to reduce the discrepancy between the ground truth measure and the output measure. However, considering that \mathcal{X} and \mathcal{Y} are supports of continuous and discrete measures respectively, it is challenging to calculate the divergence between these two heterogeneous measures which have few overlaps. Therefore, the optimization will become laborious and sometimes impractical. To address this limitation, we seek to propose a new method which can efficiently compute the distance.

3.2 Wasserstein and Sinkhorn Divergences

Wasserstein distance has recently been used as an efficient way to associate different measures. It aims to calculate the minimum discrepancy by finding the optimal transport map $\pi(x, y)$, which describes the amount of mass transporting from output measure to ground truth measure. Using $c(x, y)$ as moving costs, Wasserstein divergence can be expressed by:

$$W(\alpha, \beta) = \min_{\pi(x, y)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y), \quad (1)$$

where $\sum_{j=1}^M \pi(x, y_j) = \alpha(x)$, $\int_{\mathcal{X}} d\pi(x, y_j) = \beta(y_j)$. However, as a linear program, this divergence suffers from a high

computational cost. One effective way to release from the burden is to find an approximating version.

Typically, entropic regularization [Cuturi, 2013] is widely used in recent papers. By defining a regularized expression, the distance changes into:

$$W_{\varepsilon}(\alpha, \beta) = \min_{\pi(x, y)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) + \varepsilon \int_{\mathcal{X} \times \mathcal{Y}} \ln \left(\frac{\pi(x, y)}{\alpha(x)\beta(y)} \right) d\pi(x, y). \quad (2)$$

Entropic parameter ε , in general, is positive, determining the degree of smoothing. When $\varepsilon \rightarrow 0$, the distance converges to unregularized one. Although the regularization helps to solve the computation problem efficiently, as ε increases, when trying to optimize α by minimizing the discrepancy, α begins to shrink and the deviation expands [Feydy *et al.*, 2019]. It is obvious when $\varepsilon \rightarrow +\infty$, α will be converged to a Dirac located at the barycenter of static β . Meanwhile, as

$$W_{\varepsilon}(\alpha, \alpha) > 0, \text{ when } \varepsilon > 0,$$

the regularized Wasserstein distance does not satisfy the identity of indiscernibles.

To address this, we extend the Wasserstein distance to Sinkhorn distance [Genevay *et al.*, 2018]:

$$S_{\varepsilon}(\alpha, \beta) = W_{\varepsilon}(\alpha, \beta) - \frac{1}{2} W_{\varepsilon}(\alpha, \alpha) - \left(\frac{1}{2} W_{\varepsilon}(\beta, \beta) \right), \quad (3)$$

where deviation is prevented by a self-correcting term $W_{\varepsilon}(\alpha, \alpha)$. This debiased formula helps to guarantee the identity of indiscernibles, $S_{\varepsilon}(\alpha, \alpha) = 0$, so that the entropic bias is eliminated when matching α to ground truth β .

Unfortunately, there is still another limitation. One major requirement of above equations is $\int_{\mathcal{X}} d\alpha(x) = \int_{\mathcal{X} \times \mathcal{Y}} \pi(x, y) = \sum_{j=1}^M \beta(y_j)$. As we cannot control the total amount of the predicted density measure, directly using this divergence will violate the requirement and trigger severe mathematical problems. Meanwhile, different from the unbalanced assumption that both measures are uncontrollable, the total amount of ground truth measure has been already known. Therefore, to adapt with our problem, we propose a novel semi-balanced form to break this constraint. Details will be explained in the following sections.

3.3 Semi-balanced Sinkhorn Divergence

Here, we give derivations and properties of our semi-balanced Sinkhorn distance. Compared to traditional distances above, it expresses the relationship between two heterogeneous measures more directly. First, we define W_{ε}^{smb} as follow:

$$W_{\varepsilon}^{smb}(\alpha, \beta) = \min_{\pi(x, y)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) + D_{\varphi} \left(\sum_{j=1}^M \pi(x, y_j) | \alpha(x) \right) + \varepsilon \int_{\mathcal{X} \times \mathcal{Y}} \ln \left(\frac{\pi(x, y)}{\alpha(x)\beta(y)} \right) d\pi(x, y), \quad (4)$$

where D_{φ} is a transport penalty which relaxes the strict constraint, allowing the amount of moving mass can be differ to measure α . From an intuitive point of view, the formula gives

α slight value changes while assigning to different points. Given \mathcal{K} is the support of measure α_1 and α_2 , the penalty is defined by:

$$D_\varphi(\alpha_1|\alpha_2) = \int_{\mathcal{K}} \varphi\left(\frac{\alpha_{1k}}{\alpha_{2k}}\right) d\alpha_{2k}. \quad (5)$$

To guarantee that D_φ is nonnegative and proper, φ function should satisfy $\varphi(1) = 0$ and be strictly convex. Typically, when $\sum_{j=1}^M \pi(x, y_j) = \alpha(x)$, $\varphi(1) = 0$, indicating that the formulation degenerates to the classical Wasserstein distance. Detailed properties and the deviation will be further explained in Section 3.5.

Considering that the self-correcting term $W_\varepsilon(\alpha, \alpha)$ is inherently satisfied the amount equivalence, we can directly keep the original balanced form. Thus, we propose the semi-balanced Sinkhorn distance as:

$$S_\varepsilon^{smb}(\alpha, \beta) = W_\varepsilon^{smb}(\alpha, \beta) - \frac{1}{2}W_\varepsilon(\alpha, \alpha) + \frac{\varepsilon^2}{2}(m(\alpha) - m(\beta))^2. \quad (6)$$

3.4 Scale Consistency at Measure Space

Most existing crowd counting datasets only provide point annotations and there is no scale information available. As a result, we cannot quantify the scale information of the instances explicitly and have to regard the ground truth as Dirac measures. This may harm the counting accuracy as the scale information has turned out to be contributive [Zhang *et al.*, 2016; Zeng *et al.*, 2017; Cao *et al.*, 2018].

Previous methods have tried to address this issue by adopting multi-scale network [Liu *et al.*, 2019a] [Sindagi and Patel, 2017] [Ma *et al.*, 2020] or designing external scale modules or losses [Xu *et al.*, 2019] [Shen *et al.*, 2018b]. Different from them, in this paper, we propose a self-supervised scale enhanced mechanism from the perspective of measure theory. Our proposed mechanism is able to measure the divergence under different scale pyramid directly without normalizing them to the same sizes.

First, We resize the original image I as $Sc(I)$, and obtain the scaled output measure $\hat{\alpha} = R_\theta(Sc(I))$. Then, we perform the same re-scale transform on $\alpha = R_\theta(I)$ to get $Sc(\alpha)$. In measure space, $\hat{\alpha}$ and $Sc(\alpha)$ are expected to be closed, and ideally, $Distance(\hat{\alpha}, Sc(\alpha)) = 0$. Otherwise, if a model is sensitive to scale variations, there will be a unignored deviation between $\hat{\alpha}$ and $Sc(\alpha)$. To improve the robustness towards scale changes, the key is to minimize the scale measure divergence. This understanding leads us to punish the deviation and design a Sinkhorn scale consistency loss based on measure theory as follows:

$$\mathcal{L}_{sc} = S_\varepsilon(\hat{\alpha}, Sc(\alpha)). \quad (7)$$

3.5 Overall Loss and Optimization

The overall loss of semi-balanced Sinkhorn with scale consistency (S3) is formulated as:

$$\mathcal{L}_{S3} = \mathcal{L}_{smb} + \lambda\mathcal{L}_{sc} = S_\varepsilon^{smb}(\alpha, \beta) + \lambda S_\varepsilon(\hat{\alpha}, \alpha'). \quad (8)$$

To compute the divergence efficiently, we will find the solution from the perspective of dual transform.

Let us first introduce the Fenchel-Legendre conjugate of φ :

$$\varphi^*(z) = \sup_x zx - \varphi(x) = -\inf_x \varphi(x) - zx. \quad (9)$$

Here we choose Kullback-Leiber as φ . The divergence functions can be written by:

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

Kantorovich [Kantorovich, 1942] proposed the dual transform of Wasserstein distance to relax the hard constraints of the deterministic nature of transportation. The dual form reads as:

$$G_\varepsilon(\alpha, \beta) = \max_{f, g} \int_{\mathcal{X}} f(x) d\alpha(x) + \sum_{j=1}^M g(y_j) \beta(y_j) = \min_{\pi(x, y)} W_\varepsilon(\alpha, \beta). \quad (11)$$

Based on the Kantorovich theorem, we can then establish the duality of semi-balanced Wasserstein divergence $W_\varepsilon^{smb}(\alpha, \beta)$ in Eq. 6 by using Lagrangian term. We derive the formula as:

$$G_\varepsilon^{smb}(\alpha, \beta) = \sup_{f, g} - \int_{\mathcal{X}} \varphi^*(-f(x)) d\alpha(x) + \sum_{j=1}^M g(y_j) \beta(y_j) - \varepsilon \int_{\mathcal{X} \times \mathcal{Y}} \varphi^*\left(\frac{f(x) + g(y) - c(x, y)}{\varepsilon}\right) d\alpha(x) d\beta(y). \quad (12)$$

While optimal solutions have been found, (i.e. $\pi(x, y)$ for the primal, $f(x)g(y)$ for the dual), there is a relationship between two forms:

$$d\pi(x, y) = \exp\left(\frac{f(x) + g(y) - c(x, y)}{\varepsilon}\right) d\alpha(x) d\beta(y).$$

Therefore, finding the optimal transport map $\pi(x, y)$ is simplified to calculating the optimal dual pair $f(x)g(y)$, which significantly reduces the number of variables and computational cost. A similar proof can be found in regularized Wasserstein [Kantorovich, 1942].

Scaling Algorithm [Sinkhorn, 1964] [Chizat *et al.*, 2018] is an efficient way to compute the optimal dual pair. It gives a positive matrix to iteratively scale the dual vectors by evaluating the primal-dual relationships.

Considering the computational cost, we are inspired by this algorithm and extend it to adjust our problem. When it comes to semi-balanced form, the first order condition of G_ε^{smb} in Eq. 12 reads as:

$$\begin{aligned} f(x)|_{g, \beta} &= -\frac{\varepsilon}{1 + \varepsilon} \ln\left(\sum_{j=1}^M \exp\left(\frac{g(y_j) - c(x, y_j)}{\varepsilon}\right) \beta(y_j)\right), \\ g(y)|_{f, \alpha} &= -\varepsilon \ln\left(\int_{\mathcal{X}} \exp\left(\frac{f(x) - c(x, y)}{\varepsilon}\right) d\alpha(x)\right). \end{aligned} \quad (13)$$

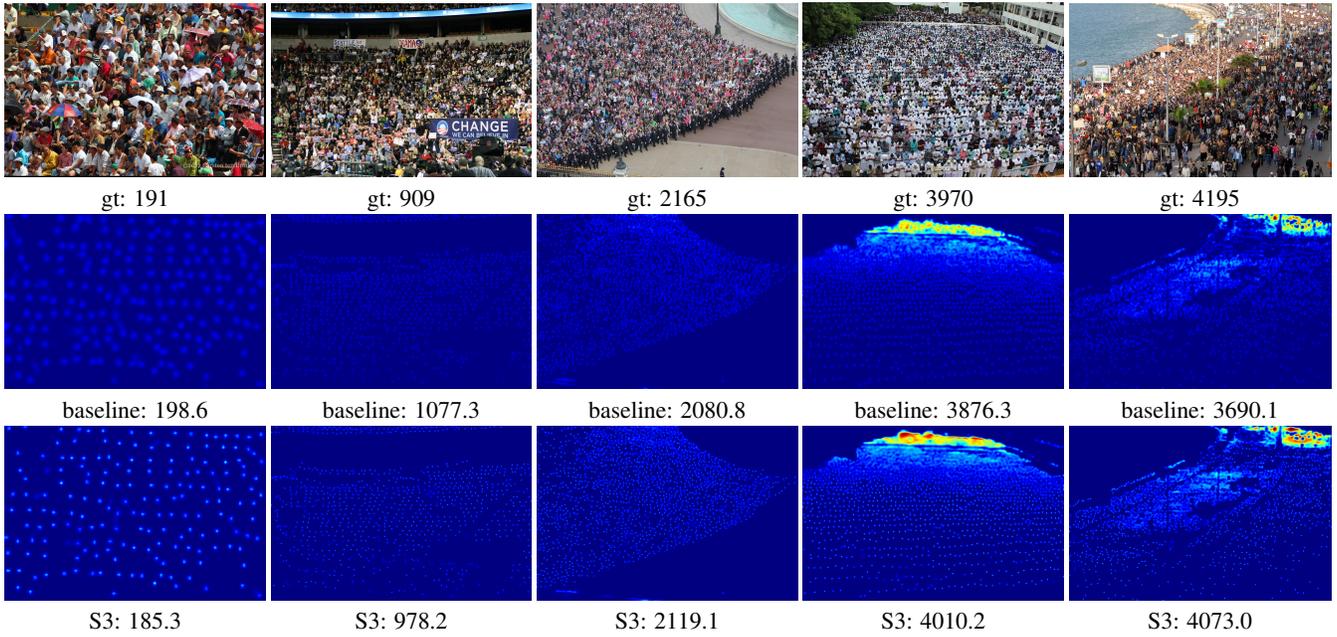


Figure 3: Visualizations of predicted density maps of the L_2 baseline and our proposed semi-balanced Sinkhorn with scale consistency. The first row: input images. The second row: predicted density maps by L_2 baseline. The third row: predicted density maps by S3.

Dataset Method	ShanghaiTech A		ShanghaiTech B		UCF-QNRF		JHU++		NWPU	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [Zhang <i>et al.</i> , 2016]	110.2	173.2	26.4	41.3	277	426	188.9	483.4	232.5	714.6
CP-CNN [Sindagi and Patel, 2017]	73.6	106.4	20.1	30.1	-	-	-	-	-	-
CSRNet [Li <i>et al.</i> , 2018]	68.2	115.0	10.6	16.0	-	-	85.9	309.2	121.3	387.8
SANet [Cao <i>et al.</i> , 2018]	67.0	104.5	8.4	13.6	-	-	91.1	320.4	190.6	491.4
CAN [Liu <i>et al.</i> , 2019a]	61.3	100.0	7.8	12.2	107.0	183.0	100.1	314.0	106.3	386.5
MBTTBF [Sindagi and Patel, 2019]	60.2	94.1	8.0	15.5	97.5	165.2	81.8	299.1	-	-
BL [Ma <i>et al.</i> , 2019]	62.8	101.8	7.7	12.7	88.7	154.8	75.0	299.9	105.4	454.2
CG-DRCN-CC [Sindagi <i>et al.</i> , 2020]	60.2	94.0	7.5	12.1	95.5	164.3	71.0	278.6	-	-
DM-Count [Wang <i>et al.</i> , 2020a]	59.7	95.7	7.4	11.8	85.6	148.3	-	-	88.4	388.6
UOT [Ma <i>et al.</i> , 2021]	58.1	95.9	6.5	10.2	83.3	142.3	60.5	252.7	87.8	387.5
L_2 Baseline	70.8	106.2	10.5	19.7	107.2	164.6	81.7	304.5	126.2	528.2
S3 (Ours)	57.0	96.0	6.3	10.6	80.6	139.8	59.4	244.0	83.5	346.9

Table 1: Comparisons with the state of the arts on ShanghaiTech, UCF-QNRF, JHU++ and NWPU four crowd benchmarks. L_2 baseline and our method are both based on VGG-19.

Here, $f(x)$ is in the *unbalanced* form (of dual vectors) while $g(y)$ is in the *balanced* one.

Thus, for balanced self-correcting term $W_\varepsilon(\alpha, \alpha)$, the symmetric optimal iteration of dual vector $p(x)$ is similar to balanced dual vector $g(y)$:

$$p(x)|_{p, \alpha} = -\varepsilon \ln \left(\int_{\mathcal{X}} \exp \left(\frac{p(x) - c(x, x)}{\varepsilon} \right) d\alpha(x) \right). \quad (14)$$

Given the density measure α and the annotated measure β , if there is no person in the image, we will only regress α to the zero vector. Otherwise, the cross correlation and self-correcting dual vectors will be computed by the Scaling Algorithms. The semi-balanced Sinkhorn distance, which is

used as counting loss \mathcal{L}_{smb} in Eq. 8, is calculated as:

$$S_\varepsilon^{smb}(\alpha, \beta) = \int_{\mathcal{X}} (-\varphi^*(-f(x)) - p(x)) d\alpha(x) + \sum_{j=1}^M g(y_j) \beta(y_j) + \frac{\varepsilon^2}{2} (m(\alpha) - m(\beta))^2. \quad (15)$$

The computation of scale consistency loss \mathcal{L}_{sc} can be performed analogously, by replacing the *unbalanced* dual vector $f(x)$ to its *balanced* form and substituting it into Eqs. 3 and 11. More details of optimization are summarized in Algorithm 1.

Algorithm 1: S3 Optimization

Input: Density regressor R with parameter θ , input image I , ground truth measure β , scale transform Sc

Output: density regressor R with optimized parameter $\hat{\theta}$

- 1 Initialize $\theta^{(1)}$;
- 2 **for** epoch $t = 1, \dots, T$ **do**
- 3 $\alpha^{(t)} = R_{\theta^{(t)}}(I)$;
- 4 $\hat{\alpha}^{(t)} = R_{\theta^{(t)}}(Sc(I))$;
- 5 Initialize $i = 1, f^{(t,i)} = 0_{N^t}, g^{(t,i)} = 0_{M^t}$;
 // $0_{N^t}, 0_{M^t}$ are zero vectors
- 6 **repeat**
- 7 $f^{(t,i+1)} = f(x)|_{g^{(t,i)}, \beta}$;
- 8 $g^{(t,i+1)} = g(y)|_{f^{(t,i+1)}, \alpha^{(t)}}$; (Eq. 13)
- 9 $i = i + 1$;
- 10 **until** convergence;
- 11 Perform iterations in steps 6-10 for $p^{(t)}$ until convergence; (Eq. 14)
- 12 $\mathcal{L}_{smb}^{(t)} = S_{smb}(f^{(t)}, g^{(t)}, p^{(t)}, \alpha^{(t)}, \beta)$; (Eq. 15)
- 13 $\mathcal{L}_{sc}^{(t)} = S_{\varepsilon}(\hat{\alpha}^{(t)}, Sc(\alpha^{(t)}))$; (Eq. 7)
- 14 $\mathcal{L}_{S3}^{(t)} = \mathcal{L}_{smb}^{(t)} + \lambda \mathcal{L}_{sc}^{(t)}$; (Eq. 8)
- 15 Minimize $\mathcal{L}_{S3}^{(t)}$ by optimizing θ ;
- 16 Update $\theta^{(t+1)}$ using Adam;
- 17 **end**
- 18 **return** $\hat{\theta} = \theta^{(t+1)}$.

4 Experimental Results

4.1 Implementation Details and Datasets

We have conducted extensive experiments on four largest crowd counting benchmarks which are widely used in recent papers. VGG-19 has been adopted as our network structure and the whole code is implemented by Pytorch. The influence of different key parameters will be detailed in Section 4.4.

ShanghaiTech [Zhang *et al.*, 2016] includes Part A and Part B. In Part A, there are 482 images with 244,167 annotated points. 300 images are divided for training and the remaining 182 images are for testing. In Part B, there are 716 images with 88,498 annotated points. 400 images are divided for training and the remaining 316 images are for testing.

UCF-QNRF [Idrees *et al.*, 2018] includes 1,535 images with 1.25 million annotated points. It has a wide range of people count and images with high resolutions. The training set contains 1,201 images and the testing set includes the rest 334 images.

JHU-CROWD++ [Sindagi *et al.*, 2020] includes 4,372 images with 1.51 million annotated points. 2,272 images are chosen for training; 500 images are for validation; and the rest 1,600 images are for testing. Compared to others, JHU-CROWD++ contains diverse scenarios and is collected under different environmental conditions of weather and illumination.

NWPU-CROWD [Wang *et al.*, 2020b] contains 5,109 images with 2.13 million annotated points. 3,109 images are divided into training set; 500 images are in validation set; and the remaining 1,500 images are in testing. Images in NWPU-CROWD are in largely various density and illumination scenes.

4.2 Comparisons with the State of the Arts

In this section, we evaluate our results on above four datasets and list ten recent state-of-the-arts methods for comparison.

The error of counting task is calculated by two commonly used metrics, Mean Absolute Error (MAE) and Mean Squared Error (MSE). The lower of both means the better performance. [Zhang *et al.*, 2016]

Visualizations of the predicted density maps are shown in Figure 3. The outputs of the proposed method appear sharp and are closed to the locations of crowds.

4.3 Quantitative Results Analysis

Counting accuracy of our method is presented in Table 1 and our proposed semi-balanced Sinkhorn with scale consistency is denoted as S3. We perform lower MAE and MSE in this task, which proves the improvements and merits of our method. Highlights are summarized as follows:

- S3 significantly improves the counting accuracy on ShanghaiTech B, UCF-QNRF, JHU++ and NWPU crowd datasets. Especially, on QNRF, S3 improves MAE and MSE values of UOT [Ma *et al.*, 2021] from 83.3 to 80.6 and from 142.3 to 139.8, respectively.
- Without any external structures or detection methods, S3 significantly improves the performance of traditional pseudo-map-regressive L_2 baseline on all four benchmarks.

4.4 Ablation Study

In this section, we hold ablation experiments to study the influence of loss terms and key parameters.

Contribution of loss terms. The overall loss is the combination of a semi-balanced Sinkhorn counting loss and a scale consistency loss. We study the contribution of each loss term in Table 2 and quantitative results verify the effectiveness of our proposed losses.

- The semi-balanced Sinkhorn counting loss \mathcal{L}_{smb} remarkably promotes traditional L_2 loss in counting accuracy. MAE and MSE are improved by **23.2** and **18.8** respectively. Then the combination of scale consistency \mathcal{L}_{sc} stabilizes our model and further reduces the error by 3.4 and 6.0.

Loss terms	L_2	\mathcal{L}_{WD}	\mathcal{L}_{smb}	S3
MAE	107.2	98.4	84.0	80.6
MSE	164.6	169.6	145.8	139.8

Table 2: Comparison of loss terms on UCF-QNRF.

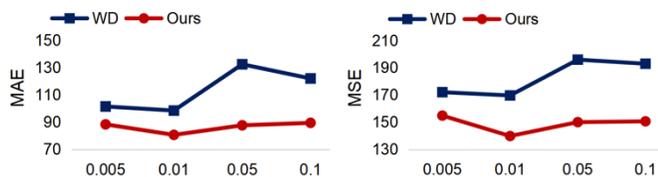


Figure 4: The influence of different smooth parameter ϵ on UCF-QNRF. Our proposed S3 performs more stable than traditional Wasserstein.

Different smooth parameter ϵ . ϵ controls the level of regularization, causing entropic bias in traditional Wasserstein distance (WD). We compare the fluctuations for using different ϵ in WD and in our proposed semi-balanced Sinkhorn with scale consistency (S3) in Figure 4.

- Compared to WD, S3 outperforms consistently and is more stable. MAE and MSE of WD vary from 98.4 to 132.3 and from 169.6 to 195.9, respectively. On contrast, those of S3 vary from 80.6 to 89.4 and from 139.8 to 150. The quantitative results can justify that our method is able to eliminate the entropic bias to some extent.

5 Conclusions

In this paper, we propose a novel measure matching based crowd counting approach, termed semi-balanced Sinkhorn with scale consistency. S3 has several advantages. 1) It avoids generating pseudo density maps with erroneous size assumptions, by allowing to use ground truth points as supervised signal. 2) The semi-balanced Sinkhorn addresses the entropic bias and amount constraints existing in other traditional measure divergences. 3) The Sinkhorn scale consistency loss stabilizes our method under the scenarios with various crowd scales. 4) The proposed pipeline works only in the learning stage and thus doesn't bring any extra computational burdens in prediction. In future, we plan to extend the proposed method to video based crowd counting.

Acknowledgements

This work is funded by National Key Research and Development Project of China under Grant No. 2019YFB1312000 and 2020AAA0105600, National Natural Science Foundation of China under Grant No. 62076195, 62006183, U20B2052, and 62006182, and by China Postdoctoral Science Foundation under Grant No. 2020M683489.

References

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *ICML*, 2017.

[Brostow and Cipolla, 2006] Gabriel J Brostow and Roberto Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *CVPR*, 2006.

[Cao *et al.*, 2018] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *ECCV*, 2018.

[Chen *et al.*, 2012] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *BMVC*, 2012.

[Chizat *et al.*, 2018] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and Francois-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 2018.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Light-speed computation of optimal transport. In *NIPS*, 2013.

[Feydy *et al.*, 2019] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Alain Trounev, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *AISTATS*, 2019.

[Genevay *et al.*, 2018] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *AISTATS*, 2018.

[He *et al.*, 2021] Yuhang He, Zhiheng Ma, Xing Wei, Xiaopeng Hong, Wei Ke, and Yihong Gong. Error-aware density isomorphism reconstruction for unsupervised cross-domain crowd counting. *AAAI*, 2021.

[Idrees *et al.*, 2018] Haroon Idrees, Muhammad 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.

[Kantorovich, 1942] Leonid Vitalievich Kantorovich. On the translocation of masses. In *Doklady Akademii Nauk*, 1942.

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

[Liu *et al.*, 2018] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G. Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *CVPR*, 2018.

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

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

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

[Ma *et al.*, 2020] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Learning scales from points: A scale-aware probabilistic model for crowd counting. In *ACM Multimedia*, 2020.

[Ma *et al.*, 2021] Zhiheng Ma, Xing Wei, Xiaopeng Hong, Hui Lin, Yunfeng Qiu, and Yihong Gong. Learning to count via unbalanced optimal transport. In *AAAI*, 2021.

[Ramdas *et al.*, 2017] Aaditya Ramdas, Trillos Nicolás García, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 2017.

- [Sajid *et al.*, 2016] Muhamad Sajid, Ali Hassan, and Shoab A Khan. Crowd counting using adaptive segmentation in a congregation. In *ICSIP*, 2016.
- [Shen *et al.*, 2018a] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. *AAAI*, 2018.
- [Shen *et al.*, 2018b] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *CVPR*, 2018.
- [Sindagi and Patel, 2017] Vishwanath A. Sindagi and Vishal M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *ICCV*, 2017.
- [Sindagi and Patel, 2019] Vishwanath A. Sindagi and Vishal M. Patel. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *ICCV*, 2019.
- [Sindagi *et al.*, 2020] Vishwanath A. Sindagi, Rajeev Yasarla, and Vishal M. Patel. JHU-CROWD++: Large-scale crowd counting dataset and a benchmark method. *arXiv preprint*, 2020.
- [Sinkhorn, 1964] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 1964.
- [Séjourné *et al.*, 2019] Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trounev, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint*, 2019.
- [Villani, 2008] Cédric Villani. *Optimal transport: old and new*. 2008.
- [Wang *et al.*, 2020a] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. *NIPS*, 2020.
- [Wang *et al.*, 2020b] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting. *arXiv preprint*, 2020.
- [Wu and Nevatia, 2005] Bo Wu and Ramakant Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, 2005.
- [Xu *et al.*, 2018] Jie Xu, Lei Luo, Cheng Deng, and Heng Huang. Multi-level metric learning via smoothed wasserstein distance. In *IJCAI*, 2018.
- [Xu *et al.*, 2019] Chenfeng Xu, Kai Qiu, Jianlong Fu, Song Bai, Yongchao Xu, and Xiang Bai. Learn to scale: Generating multipolar normalized density maps for crowd counting. In *ICCV*, 2019.
- [Yan *et al.*, 2019] Zhaoyi Yan, Yuchen Yuan, Wangmeng Zuo, Xiao Tan, Yezhen Wang, Shilei Wen, and Errui Ding. Perspective-guided convolution networks for crowd counting. In *ICCV*, 2019.
- [Zeng *et al.*, 2017] Lingke Zeng, Xiangmin Xu, Bolun Cai, Suo Qiu, and Tong Zhang. Multi-scale convolutional neural networks for crowd counting. In *ICIP*, 2017.
- [Zhang *et al.*, 2016] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, 2016.
- [Zhang *et al.*, 2018] Lu Zhang, Miaoqing Shi, and Qiaobo Chen. Crowd counting via scale-adaptive convolutional neural network. In *WACV*, 2018.
- [Zhang *et al.*, 2019] Anran Zhang, Jiayi Shen, Zehao Xiao, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Relational attention network for crowd counting. In *ICCV*, 2019.