PARTIAL WASSERSTEIN ADVERSARIAL NETWORK FOR NON-RIGID POINT SET REGISTRATION

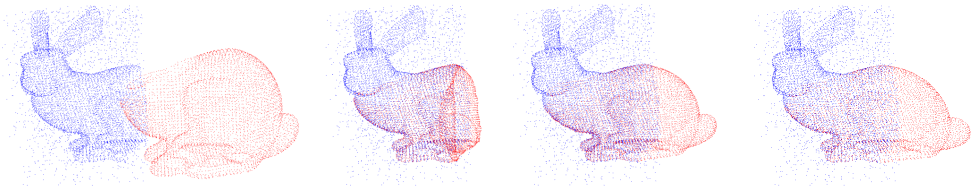
Anonymous authors
Paper under double-blind review

ABSTRACT

Given two point sets, the problem of registration is to recover a transformation that matches one set to the other. This task is challenging due to the presence of a large number of outliers, the unknown non-rigid deformations and the large sizes of point sets. To obtain strong robustness against outliers, we formulate the registration problem as a partial distribution matching (PDM) problem, where point sets are regarded as discrete distributions, and the goal is to partially match them. To handle large point sets, we propose a method for large scale PDM problem by utilizing the partial Wasserstein-1 (PW) discrepancy, which we show can be efficiently optimized. Specifically, we theoretically derive the Kantorovich–Rubinstein duality for the PW discrepancy, and show its gradient can be explicitly computed. Based on these theoretical results, we propose a partial Wasserstein adversarial network (PWAN), which approximates the PW discrepancy by a neural network, and learns the transformation adversarially with the network. It also incorporates an efficient coherence regularizer for non-rigid transformations to avoid unrealistic deformations. We evaluate PWAN on practical point set registration tasks, and show that the proposed PWAN is robust, scalable and performs more favorably than the state-of-the-art methods.

1 INTRODUCTION

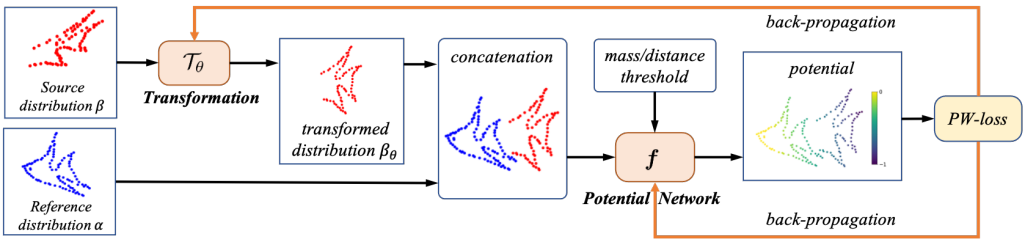
Point set registration is a fundamental task in many computer vision applications, such as object tracking (Gao & Tedrake, 2018), shape retrieval (Berger et al., 2017) and contour matching (Avots et al., 2019). As illustrated in Fig. 1, given two point sets, *i.e.*, a reference set and a source set, representing two partially overlapped shapes, the goal of this task is to recover an appropriate transformation that matches the source set to the reference one. It is challenging due to the following factors: 1) The point sets may contain a fraction of outliers which do not have valid correspondences in the other point set, such as the noise points and the points in the non-overlapped region of the shapes. 2) The number of points consisted in the point sets may be huge. 3) The deformation of point sets can be non-rigid.



(a) Input sets (b) Result of DM (c) Result of PDM (d) Ground truth

Figure 1: An example of non-rigid point set registration using the distribution matching (DM) formulation (Hirose, 2021a) and the proposed partial distribution matching (PDM) formulation.

The registration problem is generally solved in the distribution matching (DM) framework, where the point sets are regarded as probability distributions, and are aligned by minimizing a discrepancy between them. To reduce the influence of outliers, practical registration methods utilize the robust discrepancies, such as Kullback-Leibler divergence (Myronenko & Song, 2010; Hirose, 2021a) and L_2 distance (Jian & Vemuri, 2011). Thus they are able to greedily align the largest possible fraction



(a) The pipeline of the proposed PWAN

(b) The convergence of the (*potential, registration*)

Figure 2: An overview of PWAN. The transformation $\mathcal{T}_\theta$ and the network $\boldsymbol{f}$ are trained adversarially.

of points while being tolerant of a small number of outliers. For point sets that are dominated by outliers, the greedy behaviors of these methods easily bias toward outliers, thus leading to degraded registration results. An example of such result is shown in Fig. 1(b).

To obtain stronger robustness against outliers, we propose to formulate the registration problem in a novel partial distribution matching (PDM) framework, where we only seek to partially match the distributions. Comparing to the DM based methods, the PDM formulation is naturally more robust against outliers. For example, in Fig. 1(c), the PDM formulation find a more natural solution where only a fraction of points are well matched.

However, existing solutions for PDM problems generally require to compute the correspondence between distributions (Chizat et al., 2018; Chapel et al., 2020), thus they are intractable for registration problems which involve large scale distributions. To handle this issue, we propose the first efficient solver for large scale PDM problem. Our method utilizes the partial Wasserstein-1 (PW) discrepancy (Figalli, 2010), which we show can be efficiently optimized without computing the correspondence. Specifically, we theoretical derived the Kantorovich–Rubinstein (KR) duality (Villani, 2003) for the PW discrepancy, and show that its gradient can be explicitly computed via its potential. Based on these results, we propose a partial Wasserstein adversarial network (PWAN), which approximates the potential by a neural network, and the unknown transformation can be trained in an adversarial way with the network. We also incorporate a coherence regularizer for the non-rigid transformation to enforce its smoothness. We note that PWAN generalizes the popular Wasserstein generative adversarial net (WGAN) (Arjovsky et al., 2017) to the PDM problem. An illustration of the proposed PWAN is presented in Fig. 2.

The contribution of this work is summarized as follows.

- Theoretically, we derive the KR duality for the mass-type PW discrepancy. We investigate the properties of mass-type PW discrepancy and the closely related distance-type PW discrepancy. In particular, we show their gradients can be efficiently computed, thus they can be optimized via gradient descent.
- Based on the theoretical results, we propose the first method, called PWAN, for large scale PDM problem. The well-known WGAN is a special case of our method in the DM setting.
- We apply PWAN to point set registration task, where point sets are regarded as discrete distributions. We experimentally show that PWAN exhibits stronger robustness against outliers than the existing methods, and can register point sets accurately even when they are dominated by outliers, such as when they contain a large fraction of noise points, or when they are only partially overlapped.

2 Related Work

There is a large body of literatures on point set registration problem (Besl & McKay, 1992; Zhang, 1994; Chui & Rangarajan, 2000; Myronenko & Song, 2010; Maiseli et al., 2017; Vongkulbhisal et al., 2018; Hirose, 2021a). The existing methods can be roughly categorized into two classes, *i.e.*, the

correspondence-based methods and the probabilistic methods. This section only discusses the latter class which is the most related to our method. More discussions can be found in Appx. F.1.

The probabilistic methods solve the registration problem as a DM problem. Specifically, most of these methods smooth the point sets as Gaussian mixture models (GMMs), and then align the distributions via minimizing the robust discrepancies between them. Coherent point drift (CPD) (Myronenko & Song, 2010) and its variants (Hirose, 2021a) are well-known methods is this class, which minimize the Kullback-Leibler (KL) divergence between the distributions. Other robust discrepancies, including the L_2 distance (Ma et al., 2013; Jian & Vemuri, 2011), the kernel density estimator (Tsin & Kanade, 2004) and the scaled Geman–McClure estimator (Zhou et al., 2016) have also been studied.

The proposed PWAN is related to the probabilistic methods because it regards point sets as distributions. However, it has two major differences from them. First, PWAN directly processes the point sets as un-normalized discrete distributions instead of smoothing them to GMMs, thus it is more concise and natural. Second, PWAN solves a PDM problem instead of the DM problem, as it only requires to match a fraction of distributions, thus it is more robust to outliers.

Some recent works have been devoted to the PDM problem using the Wasserstein-type discrepancy. Bonneel & Coeurjolly (2019) proposed a fast algorithm based on the sliced Wasserstein metric for a special PDM problem, where a small distribution is fully matched to the large one. Various partial or unbalanced discrepancies (Chizat et al., 2018; Séjourné et al., 2019; Chapel et al., 2020) have been utilized for the PDM problem, however, they are generally not scalable to large distributions as they require to compute the correspondence between distributions. The distance-type PW discrepancy used in our paper has been utilized for imaging problem (Lellmann et al., 2014; Schmitzer & Wirth, 2019), but these methods do not directly apply to distributions in continuous space.

3 Preliminaries on Partial Wasserstein-1 Discrepancies

This section introduces the major tools used in this work, *i.e.*, two types of PW discrepancies: the partial-mass Wasserstein-1 discrepancy (Figalli, 2010; Caffarelli & McCann, 2010) and the bounded-distance Wasserstein-1 discrepancy (Lellmann et al., 2014; Bogachev, 2007). For simplicity, we call them the *mass-type* and the *distance-type* PW discrepancy respectively.

Given two discrete distributions $\alpha = \sum_{x_i \in X} a_i \delta_{x_i}$ and $\beta = \sum_{y_j \in Y} b_j \delta_{y_j}$, where point sets $X = \{x_i\}_{i=1}^q$ and $Y = \{y_j\}_{j=1}^r$ are in a compact metric space Ω, δ is the Dirac function, and $a = [a_i]_{i=1}^q \in \mathbb{R}_+^q$ and $b = [b_j]_{j=1}^r \in \mathbb{R}_+^r$ are the associated mass vectors, the mass-type PW discrepancy (Figalli, 2010) between α and β is defined by the optimal transport problem which seeks to transports at least $m \leq \min(m_\alpha, m_\beta)$ unit mass (where $m_\alpha = \sum_i a_i$ and $m_\beta = \sum_j b_j$) from α to β in the cheapest way. Formally, we denote $\mathcal{L}_{M,m}$ the mass-type PW discrepancy with mass threshold m, and define the *primal form* of $\mathcal{L}_{M,m}(\alpha, \beta)$ as

$$\mathcal{L}_{M,m}(\alpha, \beta) = \min_{\mathbf{P} \in \Gamma_m(\alpha,\beta)} \sum_{i,j} \mathbf{P}_{i,j} \, \boldsymbol{d}(x_i, y_j), \tag{1}$$

where $\mathbf{P}_{i,j}$ encodes the amount of mass transported from x_i to y_j, and the feasible set $\Gamma_m(\alpha, \beta)$ is given by $\Gamma_m(\alpha, \beta) = \{\mathbf{P} \in \mathbb{R}_+^{q \times r} | \ \mathbf{P}\mathbb{1}_r \leq a, \ \mathbf{P}^T \mathbb{1}_q \leq b, \ \mathbb{1}_q^T \mathbf{P} \mathbb{1}_r \geq m\}$.

We also consider the optimal transport problem with bounded transport distance (Lellmann et al., 2014). Formally, we denote $\mathcal{L}_{D,h}$ the distance-type PW discrepancy with a transport distance threshold h, and define the primal form of $\mathcal{L}_{D,h}$ as the Lagrangian of problem (1), where the mass constraint is soften by a Lagrange multiplier $h > 0$:

$$\mathcal{L}_{D,h}(\alpha, \beta) = \min_{\mathbf{P} \in \Gamma_0(\alpha,\beta)} \sum_{i,j} \mathbf{P}_{i,j} \, \boldsymbol{d}(x_i, y_j) - h \sum_{i,j} \mathbf{P}_{i,j}. \tag{2}$$

In this work, we use $\mathcal{L}_{D,h}$ and $\mathcal{L}_{M,m}$ as two types of objective functions to partially match the distributions. Note these two objective functions are closely related. Because according to the Lagrange duality, for a fixed (α, β, m), there exists a $h^* \geq 0$, such that the solution to $\mathcal{L}_{D,h^*}(\alpha, \beta)$ recovers to that to $\mathcal{L}_{M,m}(\alpha, \beta)$. Nevertheless, these two objective functions are not equivalent. Because when α or β varies during iterations, the corresponding h^* varies accordingly. In other words, minimizing $\mathcal{L}_{M,m}$ is generally not equivalent to minimizing $\mathcal{L}_{D,h}$ with any fixed h. We present more detailed preliminaries in Appx. A.

4 THE PROPOSED APPROACH

In this section, we present the details of the proposed PWAN. We first formulate the problem in Sec. 4.1. Then we present an efficient method to solve the problem in Sec. 4.2 and 4.3. We finally summarize our algorithm in Sec. 4.4.

4.1 PROBLEM FORMULATION

Let $Y = \{y_j\}_{j=1}^r \subseteq \Omega$ and $X = \{x_i\}_{i=1}^q \subseteq \Omega$ be the source and reference point sets, where $\Omega \subseteq \mathbb{R}^3$ is a closed bounding box of the points. We seek to solve the registration problem in a PDM formulation. Formally, we define the corresponding reference and source distribution as $\alpha = \sum_{x_i \in X} a_i \delta_{x_i}$ and $\beta = \sum_{y_j \in Y} b_j \delta_{y_j}$ respectively, where $a_i, b_j \in \mathbf{R}_+$ are the mass assigned to each point and are fixed to 1 in this paper. Let $\mathcal{T}_\theta : \Omega \to \Omega$ denote a non-rigid transformation parametrized by θ and $\beta_\theta = \sum_{y_j \in Y} b_j \delta_{\mathcal{T}_\theta(y_j)}$ denote the transformed β. Our goal is to find an optimal θ so that β_θ is partially matched to α, *i.e.*, we seek to solve

$$\min_\theta \mathcal{L}(\alpha, \beta_\theta) + \mathcal{C}(\mathcal{T}_\theta), \tag{3}$$

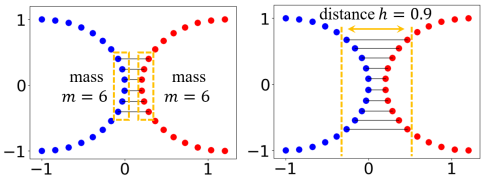
where $\mathcal{L}$ is $\mathcal{L}_{M,m}$ (1) or $\mathcal{L}_{D,h}$ (2), which measures the closeness between the nearest fraction of α and β_θ, and the coherence energy $\mathcal{C}$ ensures $\mathcal{T}_\theta$ is sufficiently smooth.

4.2 REGISTRATION WITH THE PW DISCREPANCIES

Using $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$ in the objective function (3) provides two natural ways to control the ratio of matching. Specifically, by expressing $\mathcal{L}_{M,m}$ or $\mathcal{L}_{D,h}$ in their respective primal form (1) or (2), we have

$$\mathcal{L}(\alpha, \beta_\theta) := \sum_{i,j} \mathbf{P}_{i,j} \boldsymbol{d}(x_i, \mathcal{T}_\theta(y_j)) + const, \tag{4}$$

where $const$ represents a constant not relevant to θ, and $\mathbf{P}$ represents the correspondence given by the solution of the primal form of $\mathcal{L}_{M,m}(\alpha, \beta_\theta)$ (1) or $\mathcal{L}_{D,h}(\alpha, \beta_\theta)$ (2). To obtain a sense of this formulation, we visualize the correspondence $\mathbf{P}$ between two C-shape point sets in Fig. 3. As can be seen, $\mathbf{P}$ establishes the correspondence between a fraction of points within the mass threshold m or the distance threshold h, and discards the other points, thus minimizing (4) is simply to match these corresponding point pairs. In other words, minimizing $\mathcal{L}_{M,m}(\alpha, \beta_\theta)$ is to match m points in β_θ that are nearest to α in the optimal way, while minimizing $\mathcal{L}_{D,h}(\alpha, \beta_\theta)$ is to match the points in β_θ within the distance h of α in the optimal way.



(a) $\mathcal{L}_{M,6}$ (b) $\mathcal{L}_{D,0.9}$

Figure 3: The computed correspondence $\mathbf{P}$.

However, for large scale point sets, optimizing (4) is computationally intractable, since it requires to solve for a matrix $\mathbf{P}$ of shape $q \times r$ in a linear programing in each iteration. Thus, it is natural to ask that how to avoid the computation of $\mathbf{P}$, so that the PW discrepancies can be optimized efficiently. This question is answered by the following two theorems, whose proofs are given in Appx. E.

Theorem 1. *$\mathcal{L}_{D,h}(\alpha, \beta)$ can be equivalently expressed as*

$$\mathcal{L}_{D,h}(\alpha, \beta) = \sup_{\substack{\boldsymbol{f} \in Lip(\Omega) \\ -h \le \boldsymbol{f} \le 0}} \sum_{x_i \in X} a_i \boldsymbol{f}(x_i) - \sum_{y_j \in Y} b_j \boldsymbol{f}(y_j) - h m_\beta. \tag{5}$$

The optimizer of problem (5) exists. Let $\boldsymbol{f}$ denote an optimizer of problem (5). When $\mathcal{T}_\theta$ is Lipschitz w.r.t. θ, $\mathcal{L}_{D,h}(\alpha, \beta_\theta)$ is continuous w.r.t. θ, and is differentiable almost everywhere. Furthermore, we have

$$\nabla_\theta \mathcal{L}_{D,h}(\alpha, \beta_\theta) = -\sum_{y_i \in Y} b_i \nabla_\theta \boldsymbol{f}(\mathcal{T}_\theta(y_i)). \tag{6}$$

Theorem 2. *$\mathcal{L}_{M,m}(\alpha,\beta)$ can be equivalently expressed as*

$$\mathcal{L}_{M,m}(\alpha,\beta) = \sup_{\substack{\boldsymbol{f}\in Lip(\Omega), h\in\mathbb{R}_+ \\ -h\leq \boldsymbol{f}\leq 0}} \sum_{x_i\in X} a_i \boldsymbol{f}(x_i) - \sum_{y_j\in Y} b_j \boldsymbol{f}(y_j) + h(m-m_\beta). \tag{7}$$

The optimizer of problem (7) exists. Let $(\boldsymbol{f},h)$ denote an optimizer of problem (7). When $\mathcal{T}_\theta$ is Lipschitz w.r.t. θ, $\mathcal{L}_{M,m}(\alpha,\beta_\theta)$ is continuous w.r.t. θ, and is differentiable almost everywhere. Furthermore, we have

$$\nabla_\theta \mathcal{L}_{M,m}(\alpha,\beta_\theta) = -\sum_{y_i\in Y} b_i \nabla_\theta \boldsymbol{f}(\mathcal{T}_\theta(y_i)). \tag{8}$$

Theorem 1 and Theorem 2 tell us two properties of $\mathcal{L}_{D,h}$ and $\mathcal{L}_{M,m}$. First, they provide the KR dual forms for $\mathcal{L}_{D,h}$ and $\mathcal{L}_{M,m}$ in euqation (5) and (7) respectively, which no longer require to compute the correspondence $\mathbf{P}$, but only require to fit a function $\boldsymbol{f}$ to $q+r$ data points instead. We call formulations (5) and (7) *KR forms*, and call the maximizers of them *potentials*. Furthermore, they claim that if $\mathcal{T}_\theta$ satisfies a mild regularity assumption, then $\mathcal{L}_{D,h}(\alpha,\beta_\theta)$ and $\mathcal{L}_{M,m}(\alpha,\beta_\theta)$ are differentiable w.r.t. θ and their gradients can be computed, which implies that they can be optimized via gradient descent. See Tab. 1 in the appendix for a comparison of equivalent formulations.

Now, we can formulate our PWAN model, which effectively optimizes the PW discrepancies, by implementing Theorem 1 and Theorem 2 into neural networks. Specifically, we compute $\mathcal{L}_{M,m}(\alpha,\beta_\theta)$ or $\mathcal{L}_{D,h}(\alpha,\beta_\theta)$ according to (7) or (5), where $\boldsymbol{f}$ is approximated by a neural network called the potential network. We note that both these two equations require $\boldsymbol{f}$ to be Lipschitz, and the output of $\boldsymbol{f}$ to be bounded in interval $[-h,0]$, where h is a fixed parameter in (5), while is a learnable variable in (7). Thus we parametrize $\boldsymbol{f}$ as $\boldsymbol{f}_{w,h} = \boldsymbol{l}(\boldsymbol{g}_w;h)$, where $\boldsymbol{g}_w$ is a neural network parametrized by w, which maps each point to a real number, and $\boldsymbol{l}(x;h) = \max\{-|x|,-h\}$ is an activate function acting on the output of $\boldsymbol{g}_w$. In this way, we ensure that $\boldsymbol{f}_{w,h}$ maps each point to a number in interval $[-h,0]$. To further ensure $\boldsymbol{f}_{w,h}$ is Lipschitz, we use an additional regularizer to penalize the gradient of its output w.r.t. its input (Gulrajani et al., 2017). Then we train $\boldsymbol{f}_{w,h}$ by gradient ascent: for $\mathcal{L}_{M,m}(\alpha,\beta_\theta)$, we update $\{w,h\}$ to maximize

$$Loss_{M,m} = \sum_{x_i\in X} a_i \boldsymbol{f}_{w,h}(x_i) - \sum_{y_j\in Y} b_j \boldsymbol{f}_{w,h}(\mathcal{T}_\theta(y_j)) + h(m-m_\beta) - \kappa \max_{\hat{x}\in X\cup\mathcal{T}_\theta(Y)} \{||\nabla_{\hat{x}} \boldsymbol{f}_{w,h}(\hat{x})||^2, 1\}, \tag{9}$$

while for $\mathcal{L}_{D,h}(\alpha,\beta_\theta)$, we update $\{w\}$ to maximize

$$Loss_{D,h} = \sum_{x_i\in X} a_i \boldsymbol{f}_{w,h}(x_i) - \sum_{y_j\in Y} b_j \boldsymbol{f}_{w,h}(\mathcal{T}_\theta(y_j)) - \kappa \max_{\hat{x}\in X\cup\mathcal{T}_\theta(Y)} \{||\nabla_{\hat{x}} \boldsymbol{f}_{w,h}(\hat{x})||^2, 1\}, \tag{10}$$

where κ is the strength of the gradient penalty and is set to $\kappa = 100$ in all our experiments. Finally, with the trained potentials network $\boldsymbol{f}_{w,h}$, the gradients $\nabla_\theta \mathcal{L}_{M,m}(\alpha,\beta_\theta)$ and $\nabla_\theta \mathcal{L}_{D,h}(\alpha,\beta_\theta)$ can be computed via back-propagating their respectively losses through θ, *i.e.*,

$$\nabla_\theta \mathcal{L}(\alpha,\beta_\theta) = -\sum_j b_j \nabla_\theta \boldsymbol{f}_{w,h}(\mathcal{T}_\theta(y_j)), \tag{11}$$

where $\mathcal{L}$ is $\mathcal{L}_{M,m}$ or $\mathcal{L}_{D,h}$. Thus $\mathcal{T}_\theta$ can be updated via gradient descent.

To provide an intuition of the KR forms, we present a simple example comparing the primal form and the learned KR form in Appx. F.2.

4.3 Optimize the Coherence Energy

This section discusses the optimization of the coherence energy, which ensures that the whole β_θ, including the points discarded by the PW discrepancy, remains smooth in the matching process. First of all, we define the non-rigid transformation $\mathcal{T}_\theta$ parametrized by $\theta = (A,t,V)$ as

$$\mathcal{T}_\theta(y_j) = y_j A + t + V_j, \tag{12}$$

where $y_j \in \mathbb{R}^{1\times 3}$ represents the coordinate of point $y_j \in Y$, $A \in \mathbb{R}^{3\times 3}$ is the linear transformation matrix, $t \in \mathbb{R}^{1\times 3}$ is the translation vector, $V \in \mathbb{R}^{r\times 3}$ is the offset vector of all points in Y, and

V_j represents the j-th row of V. According to this definition, $\mathcal{T}_\theta$ is the composition of two simple transformations: a linear transformation and the adding of an offset vector to each point. Despite its simplicity, $\mathcal{T}_\theta$ includes several useful transformations as its special case. For example, when $V = 0$ and $A \in SO(3)$, $\mathcal{T}_\theta$ becomes the rigid transformation, and when $A = I$ and $t = 0$, $\mathcal{T}_\theta$ becomes the "drift" transformation in Myronenko & Song (2010). Note that $\mathcal{T}_\theta$ is Lipschitz w.r.t. θ (Proposition 27 in the appendix), thus it satisfies the requirement of Theorem 1 and 2.

Now we define the coherence energy similar to Myronenko & Song (2010); Hirose (2021a), *i.e.*, we enforce that V varies smoothly in space. Formally, let $\mathbf{G} \in \mathbb{R}^{r\times r}$ be a kernel matrix, *e.g.*, the Gaussian kernel $\mathbf{G}_\rho(i,j) = e^{-||y_i - y_j||^2/\rho}$, and σ be a positive number. The coherence energy is defined as

$$\mathcal{C}(\mathcal{T}_\theta) = \lambda Tr(V^T(\sigma\mathcal{I} + \mathbf{G}_\rho)^{-1}V), \tag{13}$$

where $\lambda \in \mathbb{R}_+$ is the strength of constraint, $\mathcal{I}$ is the identity matrix, and $Tr(\cdot)$ is the trace of a matrix.

Since the matrix inversion $(\sigma\mathcal{I} + \mathbf{G}_\rho)^{-1}$ is computationally intractable for large scale point sets, we approximate it via the Nyström method (Williams & Seeger, 2000), and obtain the gradient

$$\frac{\partial \mathcal{C}(\mathcal{T}_\theta)}{\partial V} \approx (2\lambda)(\sigma^{-1}V - \sigma^{-2}\mathbf{Q}(\Lambda^{-1} + \sigma^{-1}\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^TV), \tag{14}$$

where $\mathbf{Q} \in \mathbb{R}^{r\times k}$, $\Lambda \in \mathbb{R}^{k\times k}$ is a diagonal matrix, and $k \ll r$. Note the computational cost of (14) is only $O(r)$ if we regard k as a constant. The detailed derivation is presented in Appx. F.3.

4.4 An adversarial algorithm

With the methods detailed in Sec. 4.2 and Sec. 4.3, we can finally solve problem (3) efficiently. Specifically, we iterate the following two steps: 1) Estimate the PW discrepancy between α and β_θ. 2) Update θ by minimizing the estimated PW discrepancies and the coherent energy. Formally, we solve the following mini-max problem by alternatively updating $\boldsymbol{f}_{w,h}$ and $\mathcal{T}_\theta$

$$\min_\theta \max_{\widetilde{w}} Loss(\alpha, \beta_\theta; \widetilde{w}) + \mathcal{C}(\mathcal{T}_\theta), \tag{15}$$

where we use $Loss_{M,m}$ (9), $\widetilde{w} = \{w, h\}$ for $\mathcal{L}_{M,m}$, or $Loss_{D,h}$ (9), $\widetilde{w} = \{w\}$ for $\mathcal{L}_{D,h}$. An illustration of our method is shown in Fig. 2. The detailed algorithm is presented in Alg. 1. For clearness, we refer to PWAN based on $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$ as mass-type PWAN (m-PWAN) or distance-type PWAN (d-PWAN) respectively.

Algorithm 1 PWAN for point set registration

Input: reference set X, source set Y, transform $\mathcal{T}_\theta$, potential network $\boldsymbol{f}_{w,h}$, network update frequency u, training type ("*mass*" or "*distance*"), mass threshold $\overline{m}$, distance threshold $\overline{h}$, training step T, coherence parameters (λ, ρ, σ), Nyström parameter k.
Output: the learned θ.

```
if training type = "mass" then
    w̃ ← (w, h);   m ← m̄;   L ← Loss_{M,m} defined in (9)
else if training type = "distance" then
    w̃ ← (w);   h ← h̄;   L ← Loss_{D,h} defined in (10)
end if
for t = 1, ..., T do
    Obtain the transformed distribution β_θ.
    for = 1, ..., u do
        Compute ∂L/∂w̃ by back-propagating L through w̃.
        Update w̃ by ascending the gradient ∂L/∂w̃.                ▷ Potential update step
    end for
    Compute ∂C/∂θ using (λ, ρ, σ, k) according to (14).
    Compute ∂L/∂θ by back-propagating (11) through θ.
    Update θ by descending the gradient ∂(L + C)/∂θ.              ▷ Registration step
end for
```

Finally, we analyse a toy example of registering the 1-dimensional distributions shown in Fig. 4. For now, we do not consider the coherent energy for simplicity. In the beginning of an iteration,

the potential $\boldsymbol{f}_{w,h}$ (shown in Fig. 4) is learned in the potential update step. Note the learned $\boldsymbol{f}_{w,h}$ may have some flat regions where it attains its maximal or minimal values (0 or $-h$) (Proposition 16 and 22 in the appendix). Thus the gradient norm of $\boldsymbol{f}_{w,h}$ is 0 in these regions. Then, the following registration step will move β_θ in the direction of $\nabla \boldsymbol{f}_{w,h}$. In other words, it will move the points outside the flat regions leftward (toward α), and it will not move the points inside the flat regions. After a sufficient number of iterations, a fraction of points in β_θ will be completely matched to α, while the other points will not be moved.

More discussions about the connections between our method and WGAN can be found in Appx. F.4.

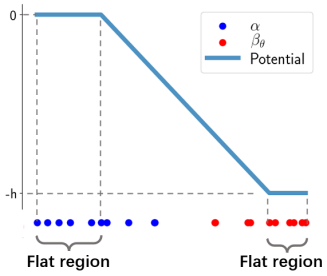


Figure 4: The learned potential for the 1-dimensional point sets.

5 Experiments and Analysis

In this section, we experimentally evaluate the proposed PWAN on point set registration. After describing the experiment settings in Sec. 5.1, we first present a toy example to highlight the robustness of the PW discrepancies in Sec. 5.2. Then we compare PWAN against the state-of-the-art methods in Sec. 5.3, and discuss its scalability in Sec. 5.4. We finally evaluate PWAN on two real datasets in Sec. 5.5. More experimental results are given in Appx. F.

5.1 Experiment Settings

As shown in Fig. 5, we use three synthesized datasets in our experiments: bunny, armadillo and monkey. The bunny and armadillo datasets are from the Stanford Repository (Standford), and the monkey dataset is from the internet. The number of points in these shape are $8,171$, $106,289$ and $7,958$ respectively. Following Hirose (2021b), we artificially deform these sets, and we evaluate the registration results via the mean square error (MSE).

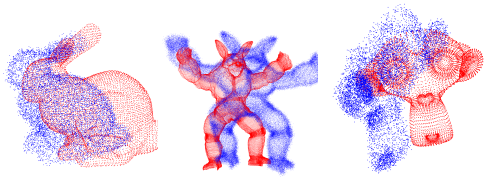


Figure 5: The datasets used in our experiments. The source point sets (blue) are synthesized by bending the reference point sets (red) in a non-linear manner.

We use a 5-layer point-wise multi-layer perception as our network (Fig. 11 in the appendix), and the parameters used in the experiments are given in Appx. F.5.

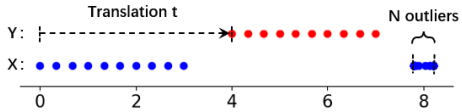
5.2 Comparison of Several Discrepancies

To provide an intuition of our PDM formulation for registration, we compare the PW discrepancy with two representative robust discrepancies used in DM based registration methods, *i.e.*, KL divergence (Hirose, 2021a) and L_2 distance (Jian & Vemuri, 2011), on a toy 1-dimensional example. For now, we do not consider the coherence energy.

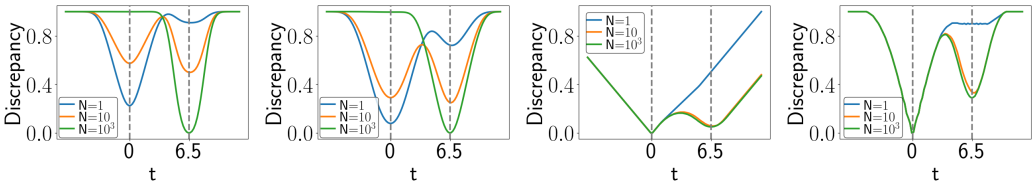
We construct the toy point sets X and Y shown in Fig. 6(a), where we first sample 10 equi-spaced data points in interval $[0, 3]$, then we define Y by translating the data points by a distance t, and define X by adding N outliers in a narrow interval $[7.8, 8.2]$ to the data points. For $N = \{1, 10, 10^3\}$, we record four discrepancies: KL, L_2, $\mathcal{L}_{M,10}$ and $\mathcal{L}_{D,2}$ between X and Y as a function of t, and present the results from Fig. 6(b) to Fig. 6(e).

As can be seen, there exist two alignment modes in this experiment, *i.e.*, $t = 0$ and $t = 6.5$, which respectively correspond to the correct alignment and the degraded alignment where Y is matched to outliers. When the number of outliers is small, *e.g.*, $N = 1$, all discrepancies admit a deep local minimum $t = 0$. However, as for KL and L_2 divergence, the local minimum $t = 0$ gradually vanishes and they gradually bias toward $t = 6.5$ as N increases. In contrast, $\mathcal{L}_{M,10}$ and $\mathcal{L}_{D,2}$ do not suffer from this issue, *i.e.*, the local minimum $t = 0$ remains deep regardless of the value of N, and the local minimum $t = 6.5$ is always shallower than the local minimum $t = 0$.

The results show a key advantage of PDM against DM for registration: The correct local minimum $t = 0$ of the DM formulation (KL and L_2) vanishes as N increases. However, thanks to the threshold of PW discrepancies, the local minimum $t = 0$ of the PDM formulation is not influenced by the remote outliers beyond the threshold, *i.e.*, it maintains a deep local minimum $t = 0$ even for large N. Thus the PDM formulation is always likely to find the correct alignment $t = 0$ even for the large N.



(a) Experiment setting of the toy example.



(b) L_2 (c) KL (d) $\mathcal{L}_{M,10}$ (e) $\mathcal{L}_{D,2}$

Figure 6: Comparison of different discrepancies on a toy point set.

5.3 Evaluation of the Registration Accuracy

We compare the performance of PWAN with four state-of-the-art methods: CPD (Myronenko & Song, 2010), GMM-REG (Jian & Vemuri, 2011), BCPD (Hirose, 2021a) and TPS-RPM (Chui & Rangarajan, 2000). We evaluate them on the following two artificial datasets.

Point sets with extra noise points. We first sample $N = 500$ random points from each of the original and the deformed sets as the source and reference sets respectively. Then we add extra uniformly distributed noise points to the reference set, and we normalize both sets to mean 0 and variance 1. We vary the number of outliers from 100 to 600 at an interval of 100, *i.e.*, the outlier/non-outlier ratio varies from 0.2 to 1.2 at an interval of 0.2.

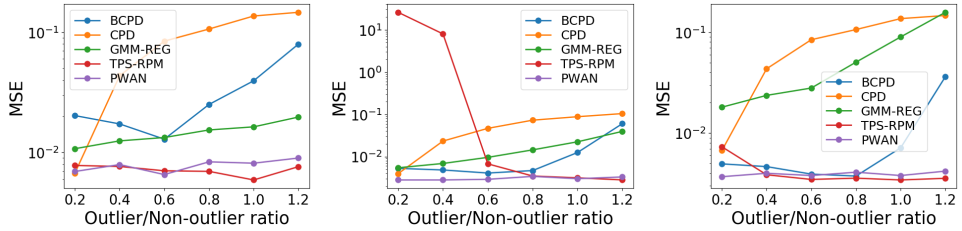
Partially overlapped point sets. We first sample $N = 1000$ random points from each of the original and the deformed sets as the source and reference sets respectively. Then for each set, we intersect it with a random plane, and we retain the points in one side of the plane and discard all points in the other side. We vary the retain ratio s from 0.7 to 1.0 at an interval of 0.05 for both the source and the reference sets, thus the minimal ratios of overlapping area are $(2s - 1)/s$ =0.57, 0.67, 0.75, 0.82, 0.89, 0.94 and 1 accordingly, and the minimal corresponding mass is $\underline{m} = (2s - 1) * 1000$.

We evaluate m-PWAN with $m = 500$ (equivalently d-PWAN with $h = +\infty$) in the first experiment, while we evaluate m-PWAN with $m = \underline{m}$ and d-PWAN with $h = 0.05$ in the second experiment. For each setting, we run all methods 100 times and report the median of MSE in Fig. 7.

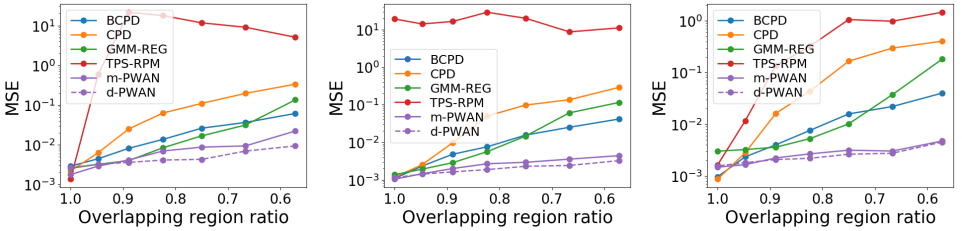
Fig. 7(a) presents the results of the first experiment. In the bunny and armadillo datasets, PWAN performs comparably with TPS-RPM, and these two methods show much lower registration errors than other methods. However, for the monkey dataset, TPS-RPM often fails to converge, while PWAN still shows low registration errors. This suggests that both TPS-RPM and PWAN are more robust against noise points than other baseline methods, while PWAN is more stable than TPS-RPM. Fig. 7(b) presents the results of the second experiment. Two types of PWANs perform comparably, and they outperform the baseline methods by a large margin when the overlap ratio is low, while perform comparably with them when the data is fully overlapped. This result suggests the proposed PWAN can effectively register the partially overlapped sets.

5.4 Evaluation of the Efficiency

To evaluate the efficiency of PWAN, we first need to investigate the influence of its parameters. In particular, the most important parameter that affects the efficiency is the network update frequency u, which controls the tradeoff between the efficiency and effectiveness. Specifically, larger u leads to more accurate estimation of the potential and the gradient, while smaller u allows for faster



(a) Evaluation of robustness against noise points.



(b) Evaluation of robustness against partial overlapping.

Figure 7: Registration accuracy of the bunny (left), monkey (middle) and armadillo (right) datasets.

estimations. We quantify the influence of u using the bunny dataset and report the results in the left panel of Fig. 8. As can be seen, both MSE and the variance of MSE decrease as u increases, while the computation time increases proportionally with u. This result suggests that more accurate and stable results can be achieved at the expense of higher computational cost, *i.e.*, larger u.

We benchmark the computation time of different methods on a computer with two Nvidia GTX TITAN GPUs and an Intel i7 CPU. We fix $u = 20$ for PWAN. We sample $q = r$ points from the bunny shape, where q varies from 10^3 to 7×10^5. PWAN is run on the GPU while the other methods are run on the CPU. We also implement a multi-GPU version of PWAN where the potential network is updated in parallel. We run each method 10 times and report the mean of the computation time in the right panel of Fig. 8. As can be seen, BCPD is the fastest method when q is small, and PWAN is comparable with BCPD when q is near 10^6. In addition, the 2-GPU version PWAN is faster than the single GPU version, and it is faster than BCPD when q is larger than 5×10^5.

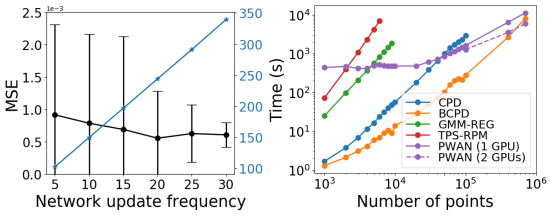


Figure 8: Scalability of our method.

5.5 Evaluation on Real Data

To demonstrate the capability of PWAN on handling point sets with non-artificial deformations, we evaluate it on a human face dataset (Zhang et al., 2008) and a 3D human dataset (DataSet). The details of this experiment are shown in Appx. F.9 and Appx. F.10.

6 Conclusion

In this paper, we formulate the point set registration task as a PDM problem, where point sets are regarded as discrete distributions and are only required to be partially matched. In order to solve the PDM problem, we develope a theory for PW discrepancy, which allows us to estimate it efficiently. Based on the theory, we proposed a PWAN method for PDM problem, and apply it to practical point sets registration task. We experimentally show that PWAN can effectively handle the point sets dominated by outliers, including those containing large fraction of noise or being partially overlapped.

There are several issues need further study. First, the computation time of PWAN is still relatively high. A possible approach to accelerate PWAN is to incorporate a forward generator network as in Sarode et al. (2019) or to use the discriminative training as in Vongkulbhisal et al. (2018). Second, it is interesting to explore PWAN in other PDM problems, such as domain adaption (Hu et al., 2020), pose estimation (Kuhnke & Ostermann, 2019) and multi-label learning (Yan & Guo, 2019).

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv: Machine Learning*, 2017.

Egils Avots, Meysam Madadi, Sergio Escalera, Jordi Gonzalez, Xavier Baro, Paul Pällin, and Gholamreza Anbarjafari. From 2d to 3d geodesic-based garment matching. *Multimedia Tools and Applications*, 78(18):25829–25853, 2019.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Matthew Berger, Andrea Tagliasacchi, Lee M. Seversky, Pierre Alliez, Gaël Guennebaud, Joshua A. Levine, Andrei Sharf, and Claudio T. Silva. A survey of surface reconstruction from point clouds. *Computer Graphics Forum*, 36(1):301–329, 2017.

P.J. Besl and H.D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

Vladimir I Bogachev. *Measure theory*, volume 1. Springer Science & Business Media, 2007.

Nicolas Bonneel and David Coeurjolly. Spot: sliced partial optimal transport. *ACM Transactions on Graphics*, 38(4):89, 2019.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

Luis A. Caffarelli and Robert J. McCann. Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of Mathematics*, 171(2):673–730, 2010.

Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Haili Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, volume 2, pp. 2044–2051, 2000.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013.

DataSet. Matching humans with different connectivity. URL http://profs.scienze.univr.it/~marin/shrec19/.

Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3483–3491, 2018.

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: asymptotic and gradient properties. In *AISTATS 2020-23nd International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 1–20, 2020.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pp. 3186–3197. PMLR, 2021.

Alessio Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010.

Andrew W. Fitzgibbon. Robust registration of 2d and 3d point sets. *Image and Vision Computing*, 21 (13):1145–1153, 2003.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Wei Gao and Russ Tedrake. Surfelwarp: Efficient non-volumetric single view dynamic reconstruction. In *Robotics: Science and Systems XIV*, volume 14, 2018.

Song Ge, Guoliang Fan, and Meng Ding. Non-rigid point set registration with global-local topology preservation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 245–251, 2014.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1574–1583. PMLR, 2019.

Sébastien Granger and Xavier Pennec. Multi-scale em-icp: A fast and robust approach for surface registration. In *ECCV '02 Proceedings of the 7th European Conference on Computer Vision-Part IV*, volume 2353, pp. 418–432, 2002.

Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *CoRR*, abs/2001.06937, 2020. URL https://arxiv.org/abs/2001.06937.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, pp. 5769–5779, 2017.

Eric Heitz, Kenneth Vanhoey, Thomas Chambon, and Laurent Belcour. A sliced wasserstein loss for neural texture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9412–9420, June 2021.

Osamu Hirose. A bayesian formulation of coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2269–2286, 2021a.

Osamu Hirose. Acceleration of non-rigid point set registration with downsampling and gaussian process regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2858–2865, 2021b.

Jian Hu, Hongya Tuo, Chao Wang, Lingfeng Qiao, Haowen Zhong, Junchi Yan, Zhongliang Jing, and Henry Leung. Discriminative partial domain adversarial network. In *European Conference on Computer Vision*, pp. 632–648, 2020.

Bing Jian and B C Vemuri. Robust point set registration using gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1633–1645, 2011.

L. V. Kantorovich. On the translocation of masses. *Journal of Mathematical Sciences*, 133(4): 1381–1382, 2006.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5258–5267, 2016.

Felix Kuhnke and Joern Ostermann. Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10164–10173, 2019.

Jan Lellmann, Dirk A. Lorenz, Carola-Bibiane Schönlieb, and Tuomo Valkonen. Imaging with kantorovich–rubinstein discrepancy. *Siam Journal on Imaging Sciences*, 7(4):2833–2859, 2014.

Jiayi Ma, Ji Zhao, Jinwen Tian, Zhuowen Tu, and Alan L. Yuille. Robust estimation of nonrigid transformation for point set registration. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2147–2154, 2013.

Baraka Maiseli, Yanfeng Gu, and Huijun Gao. Recent developments and trends in point set registration methods. *Journal of Visual Communication and Image Representation*, 46(46):95–106, 2017.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2): 583–601, 2002.

Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12):2262–2275, 2010.

Vinit Sarode, Xueqian Li, Hunter Goforth, Yasuhiro Aoki, Rangaprasad Arun Srivatsan, Simon Lucey, and Howie Choset. Pcrnet: Point cloud registration network using pointnet encoding, 2019.

Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019.

Bernhard Schmitzer and Benedikt Wirth. A framework for wasserstein-1-type metrics. *Journal of Convex Analysis*, 26(2):353–396, 2019.

Standford. The stanford 3d scanning repository. *[Online]. Available: http://graphics.stanford.edu/data/3Dscanrep/.*

C.V. Stewart, Chia-Ling Tsai, and B. Roysam. The dual-bootstrap iterative closest point algorithm with application to retinal image registration. *IEEE Transactions on Medical Imaging*, 22(11): 1379–1394, 2003.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

Yanghai Tsin and Takeo Kanade. A correlation-based approach to robust point set registration. In *European Conference on Computer Vision*, pp. 558–569, 2004.

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Jayakorn Vongkulbhisal, Benat Irastorza Ugalde, Fernando De la Torre, and Joao P. Costeira. Inverse composition discriminative optimization for point cloud registration. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2993–3001, 2018.

Christopher K. I. Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, volume 13, pp. 682–688, 2000.

Yan Yan and Yuhong Guo. Adversarial partial multi-label learning. *arXiv preprint arXiv:1909.06717*, 2019.

Li Zhang, Noah Snavely, Brian Curless, and Steven M Seitz. Spacetime faces: High-resolution capture for modeling and animation. In *Data-Driven 3D Facial Animation*, pp. 248–276. Springer, 2008.

Zhengyou Zhang. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, 13(2):119–152, 1994.

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *European Conference on Computer Vision*, pp. 766–782, 2016.

APPENDIX

In this appendix, we provide theoretical justifications of our algorithm and more experimental results.

The appendix is organized as follows. We first present a more general introduction of the optimal transport problem and fix the notations in Sec. A. Then we introduce existing computational approach of optimal transport problem in Sec. B. We derive our formulation in Sec. C. We further discuss the properties of our formulations in Sec. D. Finally, we prove the theorems in the main text in Sec. E and we present more details for the main text in Sec F.

A PRELIMINARIES

This section introduces the optimal transport (OT) problem and fix the notations for later sections.

A.1 OPTIMAL TRANSPORT

OT is a classic problem dating back to Monge and Kantorovich. It requires to solve the following transportation problem: let α be the distribution of warehouses storing some raw materials, and β be the distribution of factories requiring these materials. Assuming the total mass of materials stored in the warehouse is m_α, and the total mass of materials required by the factories is m_β. How to transport at least m ($m \leq \min(m_\alpha, m_\beta)$) mass materials from α to β so that the total cost is minimized?

Formally, let Ω be compact metric space and $\mathcal{M}_+(\Omega)$ be the set of non-negative measures defined on Ω. Define the source and target measures $\alpha, \beta \in \mathcal{M}_+(\Omega)$, and a continuous cost function $c : \Omega \times \Omega \to \mathbb{R}^+$. OT seeks to solve the following optimization problem

$$\mathcal{L}_{M,m}(\alpha, \beta) = \inf_{\pi \in \Gamma_m(\alpha,\beta)} \int_{\Omega\times\Omega} c(x, y) d\pi(x, y), \tag{16}$$

where $\Gamma_m(\alpha, \beta)$ are the set of non-negative measures π defined on $\Omega \times \Omega$ satisfying

$$\pi(A \times \Omega) \leq \alpha(A), \quad \pi(\Omega \times A) \leq \beta(A) \quad and \quad \pi(\Omega \times \Omega) \geq m$$

for all measurable set $A \subseteq \Omega$. For ease of notations, we abbreviate $m_\alpha = \alpha(\Omega)$, $m_\beta = \beta(\Omega)$ and $m(\pi) = \pi(\Omega \times \Omega)$.

In the complete OT problem, it is generally assumed that $m_\alpha = m_\beta = m$, *i.e.*, the source and the target distributions contain equal mass of materials, and *all* materials should be transported. However, the more general partial optimal transport (POT) problem (Figalli, 2010; Caffarelli & McCann, 2010) only requires $m \leq min(m_\alpha, m_\beta)$, *i.e.*, the source and target distributions may contain different mass of materials, and only a *partial* of mass is required to transported.

A.2 PARTIAL WASSERSTEIN-1 DISCREPANCY

In this paper, we focus on a specific POT problem where the cost function is a distance in the metric space Ω, *i.e.*, $c(x, y) = d(x, y)$, where d is the distance function defined in Ω. In this case, we called $\mathcal{L}_{M,m}$ the mass-type partial Wasserstein-1 (PW) discrepancy.

In the complete OT case, this type of OT problem defines the Wasserstein-1 metric, which is also known as the earth move distance (EMD) between distributions. According to the Kantorovich–Rubinstein duality (Kantorovich, 2006), the Wasserstein-1 metric can be equivalently expressed as

$$\mathcal{W}_1(\alpha, \beta) = \sup_{\boldsymbol{f} \in Lip(\Omega)} \int_\Omega \boldsymbol{f} d\alpha - \int \boldsymbol{f} d\beta. \tag{17}$$

Note that this formulation offers a more efficient implement of Wasserstein-1 metric than the primal form 16, as it only requires to solve for a function with a local constraint in Ω instead of a transport plan with global constraints in $\Omega \times \Omega$.

Several methods have been proposed to generalize (17) to the unbalanced OT (Chizat et al., 2018; Schmitzer, 2019), *i.e.*, OT problems with extra regularizers. See Schmitzer & Wirth (2019) for an unified framework of this type of generalizations. Amongst these generalizations, KR metric (Lellmann

et al., 2014) is closely related to the POT problem 16 considered in this paper, as it can be regarded as a Lagrangian of problem 16 where the mass constraint is soften. Specifically, the primal form KR metric with parameter $h > 0$ is defined as

$$KR_h(\alpha, \beta) = \inf_{\pi \in \Gamma_0(\alpha,\beta)} \int_{\Omega \times \Omega} d(x, y) d\pi(x, y) - hm(\pi) + \frac{h}{2}(m_\alpha + m_\beta). \tag{18}$$

Note the KR metric has several equivalent primal forms, such as the unbalanced optimal transport formulation with total-variation regularizer (Schmitzer & Wirth, 2019), or the flow-based formulations in Lellmann et al. (2014). It is important to notice that the KR metric has a natural explanation that it requires to find a optimal plan whose *transport distance does not exceed* h. This can be seen by re-writing problem 18 as $KR_h(\alpha, \beta) = \inf \int (d(x, y) - h) d\pi(x, y) + const$, and noticing that if $d(x, y) > h$, the solution π^* to problem 18 should satisfy $\pi^*(x, y) = 0$, because otherwise π^* can be further improved by setting $\pi^*(x, y)$ to 0.

Importantly, the KR metric is known to have an equivalent form (Lellmann et al., 2014; Schmitzer & Wirth, 2019)

$$KR_h(\alpha, \beta) = \sup_{\substack{\boldsymbol{f} \in Lip(\Omega) \\ |\boldsymbol{f}| \leq \frac{h}{2}}} \int_\Omega \boldsymbol{f} d\alpha - \int \boldsymbol{f} d\beta, \tag{19}$$

which is very similar to (17), since they both involve the maximization of a Lipschitz function. We called this type of formulations *KR form*s, and the solution to KR forms *potential*s. We refer readers to Villani (2009) for more historical contents.

A.3 Theoretical contributions

The contributions of this work are summarized as follows.

- We present the exact KR form of $\mathcal{L}_{M,m}$ in Proposition. 3.
- We investigate some properties of the KR form of $\mathcal{L}_{M,m}$. In particular, we show that the potentials of $\mathcal{L}_{M,m}$ exist in Sec. D.1, which is the foundation of the gradient computation in KR form in Sec. D.5. Furthermore, some properties of the potentials are proved in Sec. D.4, including the "flatness" of the potentials, which justifies the ability of our method to discard a fraction of distributions. We also present the properties of $\mathcal{L}_{D,h}$ as analogues of those of $\mathcal{L}_{M,m}$.
- We present the connections between various related formulations in Sec. D.2 and D.3, including the primal, dual and KR forms of $\mathcal{L}_{M,m}$, $\mathcal{L}_{D,h}$ and $\mathcal{W}_1$. These results connect our method and WGAN, and they generalize the classic OT theories. In particular, the $\overline{c(h)}$-transform connecting the dual and the KR form is the partial version of the classic c-transform, and the resulting Proposition 9 and 7 are analogues of the property of Kantorovich duality (Theorem 5.9 (c) in Villani (2009)).

A.4 Notations

- (Ω, d): a metric d associated to a compact metric space Ω. For example, Ω can be a closed cubic in $\mathbb{R}^3$ and d can be the Euclidean distance.
- $C(\Omega)$: the set of continuous bounded function defined on Ω equipped with the supreme norm.
- $Lip(\Omega) \subseteq C(\Omega)$: the set of 1-Lipschitz function defined on Ω.
- $\mathcal{M}(X)$: the space of Radon measures on space X.
- $\pi^1_\#$: The marginal of π on its first variable. Similarly, $\pi^2_\#$ represent the marginal of π on its second variable.
- Given a function $\mathbf{F}$: $X \to \mathbb{R} \cup +\infty$, the Fenchel conjugate of $\mathbf{F}$ is denoted as $\mathbf{F}^*$ and is given by:

$$\mathbf{F}^*(x^*) = \sup_{x \in X} < x, x^* > -\mathbf{F}(x), \quad \forall x^* \in X^* \tag{20}$$

  where X^* is the dual space of X and $< \cdot >$ is the dual pairing.

B EXISTING COMPUTATIONAL APPROACHES

The computation of OT problem is an active field in machine learning. In this section, we briefly discuss three major classes of approaches and relate our method to the existing ones.

One class of the most well-developed OT solver is based on the entropic regularizer. This class of approaches relax the primal problem by adding an entropy regularizer to the transport plan, then solve the relaxed problem via the Sinkhorn algorithm (Cuturi, 2013). However, the direct application of the algorithm in Cuturi (2013) has two drawbacks. First, the entropic bias introduced by the regularizer always leads to undesired behaviors. Second, the computational cost is high for large scale problems, since the Sinkhorn algorithm iteratively updates the whole transport plan. Some works have been devoted to address these two issues. To get rid of the bias, Genevay et al. (2018; 2019) proposed the Sinkhorn divergence as an unbiased version of the entropic OT. To improve the efficiency, Schmitzer (2019) proposed some acceleration techniques such as muti-scale computing, and Fatras et al. (2020) proposed to consider the mini-batch OT problem to avoid the computation of the complete transport plan. The generalization of this type of approaches to unbalanced or partial OT problem was studied in Benamou et al. (2015); Chizat et al. (2018); Séjourné et al. (2019); Fatras et al. (2021).

The second class of approaches solve the sliced OT problem (Bonneel et al., 2015; Kolouri et al., 2016). They avoid computing OT problems in high dimensional space by projecting the distributions onto random 1-dimensional lines, and solving a 1-dimensional OT problem each time. Due to their simplicity and efficiency, this class of approaches have been applied to several fields in computer vision, such as generative modelling (Deshpande et al., 2018) and texture synthesis (Heitz et al., 2021). Recently, Bonneel & Coeurjolly (2019) proposed an algorithm for a special case of partial sliced OT problem, where a small distribution is completely matched to a fraction of a large distribution. However, this method can not handle the general partial sliced OT problem where both distributions are not completely overlapped.

Our method belongs to the third class of approaches which focus on a specific type of OT problem: the Wasserstein-1 type problem. The foundation of this type of approach is the Kantorovich–Rubinstein duality (17), which claims that the Wasserstein-1 distance can be efficiently computed by estimating a Lipschitz potential function. This property was directly exploited in the popular Wasserstein GAN model (Arjovsky et al., 2017). Some works (Lellmann et al., 2014; Schmitzer & Wirth, 2019) generalized this duality to the unbalanced Wasserstein-1 type problem and applied them to imaging problems, but the exact partial Wasserstein-1 problem ((16) with $c = d$) has not been considered. Our method completes these approaches in a sense that we solve the exact partial Wasserstein-1 problem. Besides, unlike the existing works (Lellmann et al., 2014; Schmitzer & Wirth, 2019) which only handle the distributions in discrete image space, our method handles distributions in the continuous space, thus it is more suitable for applications in machine learning such as point sets registration.

C OUR FORMULATIONS

In this section, we first define the distance-type PW discrepancy by reformulating the KR metric. Then we derive a KR form for the mass-type PW discrepancy 16 by utilizing the equivalence of the primal and the dual form of KR metric.

C.1 DISTANCE-TYPE PW DISCREPANCY

First of all, we consider the Lagrangian POT problem by introducing a Lagrange multiplier $h > 0$ conjugate to the minimal mass constraint in problem 16,

$$\mathcal{L}_{D,h}(\alpha, \beta) = \inf_{\pi \in \Gamma_0(\alpha,\beta)} \int_{\Omega \times \Omega} c(x, y) d\pi(x, y) - hm(\pi). \tag{21}$$

When $c(x, y) = d(x, y)$, we immediately obtain the KR form of this problem by reformulating the KR metric 18. This is stated in the following proposition.

Proposition 1 (KR form of $\mathcal{L}_{D,h}$)**.** *When $c(x, y) = d(x, y)$, problem 21 can be reformulated as*

$$\mathcal{L}_{D,h}(\alpha, \beta) = \sup_{\substack{\boldsymbol{f} \in Lip(\Omega) \\ -h \leq \boldsymbol{f} \leq 0}} \int_{\Omega} \boldsymbol{f} d\alpha - \int_{\Omega} \boldsymbol{f} d\beta - hm_{\beta}. \tag{22}$$

Proof. This is a direct result of the equivalence of (19) and (18). □

As mentioned earlier, $\mathcal{L}_{D,h}$ has a natural explanation that it requires to find a optimal plan within the distance h, we thus called the formulation 22 distance-type PW discrepancy. We note that the last term $-hm_\beta$ is a constant term, thus can be omitted if one only wishes to compute the potential $\boldsymbol{f}$.

We define a functional associated to (22) as follows.

Definition 1 ($\mathbf{L}_{D,h}$)**.** *Define a functional* $\mathbf{L}_{D,h}$ *associated to* (22) *as*

$$\mathbf{L}_{D,h}^{\alpha,\beta}(\boldsymbol{f}) = \begin{cases} \int_\Omega \boldsymbol{f} d\alpha - \int_\Omega \boldsymbol{f} d\beta - hm_\beta & \boldsymbol{f} \in Lip(\Omega)\ ,\ -h \le \boldsymbol{f} \le 0, \\ -\infty & else \end{cases}$$

With this definition, problem 22 becomes

$$\mathcal{L}_{D,h}(\alpha,\beta) = \sup_{\boldsymbol{f}\in C(\Omega)} \mathbf{L}_{D,h}^{\alpha,\beta}(\boldsymbol{f})$$

Finally we remark that the maximizer of $\mathbf{L}_{D,h}^{\alpha,\beta}$ exists.

Proposition 2 (Existence of optimizer Schmitzer & Wirth (2019))**.** *For* $\alpha, \beta \in \mathcal{M}_+(\Omega)$ *and* $h > 0$*, there exists* $\boldsymbol{f} \in C(\Omega)$ *such that* $\mathcal{L}_{D,h}(\alpha,\beta) = \mathbf{L}_{D,h}^{\alpha,\beta}(\boldsymbol{f})$.

Proof. This proposition can be seen as a special case of Proposition 2.11 in Schmitzer & Wirth (2019). □

C.2 Mass-type PW discrepancy

We now derive the KR form of $\mathcal{L}_{M,m}(\alpha,\beta)$. To this end, we first derive the dual forms of $\mathcal{L}_{M,m}(\alpha,\beta)$ and $\mathcal{L}_{D,h}(\alpha,\beta)$, which reveal a simple relation between these two problems. By incorporating the KR form 22 of $\mathcal{L}_{D,h}(\alpha,\beta)$ into this relation, we immediately obtain the KR form of $\mathcal{L}_{M,m}(\alpha,\beta)$.

To begin with, we present the dual form of $\mathcal{L}_{M,m}(\alpha,\beta)$ and $\mathcal{L}_{D,h}(\alpha,\beta)$ in the following two propositions.

Proposition 3 (Dual form of $\mathcal{L}_{M,m}$)**.** *Problem 16 can be equivalently expressed as*

$$\mathcal{L}_{M,m}(\alpha,\beta) = \sup_{(\boldsymbol{f},\boldsymbol{g},h)\in\mathbf{R}} \int_\Omega \boldsymbol{f} d\alpha + \int_\Omega \boldsymbol{g} d\beta + mh. \tag{23}$$

where the feasible set $\mathbf{R}$ *is*

$$\mathbf{R} = \Big\{(\boldsymbol{f},\boldsymbol{g},h) \in C(\Omega)\times C(\Omega)\times \mathbb{R}_+ | \boldsymbol{f} \le 0,\ \boldsymbol{g} \le 0,\ c(x,y) - h - \boldsymbol{f}(x) - \boldsymbol{g}(y) \ge 0, \forall x,y \in \Omega\Big\} \tag{24}$$

In addition, the infimum in Problem 16 is attained.

Proof. We prove this proposition via Fenchel-Rockafellar duality. We first define space E: $C(\Omega) \times C(\Omega) \times \mathbb{R}$, space F: $C(\Omega \times \Omega)$, and a linear operator $\mathcal{A}$: $E \to F$ as

$$\mathcal{A}(\boldsymbol{f},\boldsymbol{g},h) : (x,y,h) \to \boldsymbol{f}(x) + \boldsymbol{g}(y) + h; \quad \forall \boldsymbol{f},\boldsymbol{g} \in C(\Omega),\ \forall h \in \mathbb{R},\ \forall x,y \in \Omega. \tag{25}$$

Then we introduce a convex function $\mathbf{H}$: $F \to \mathbb{R} \cup +\infty$ as

$$\mathbf{H}(u) = \begin{cases} 0 & if\ u \ge -c \\ +\infty & else \end{cases} \tag{26}$$

and $\mathbf{L}$: $E \to \mathbb{R} \cup +\infty$ as

$$\mathbf{L}(\boldsymbol{f},\boldsymbol{g},h) = \begin{cases} \int \boldsymbol{f} d\alpha + \int \boldsymbol{g} d\beta + hm & if\ \boldsymbol{f} \ge 0,\ \boldsymbol{g} \ge 0,\ h \le 0 \\ +\infty & else \end{cases} \tag{27}$$

We can check when $\boldsymbol{f} \equiv \boldsymbol{g} \equiv 1$ and $h = -1$, $\mathbf{H}$ is continuous at $\mathcal{A}(\boldsymbol{f},\boldsymbol{g},h)$. Thus by Fenchel-Rockafellar duality, we have

$$\inf_{(\boldsymbol{f},\boldsymbol{g},h)\in E} \mathbf{H}(\mathcal{A}(\boldsymbol{f},\boldsymbol{g},h)) + \mathbf{L}(\boldsymbol{f},\boldsymbol{g},h) = \sup_{\pi\in\mathcal{M}(\Omega\times\Omega)} -\mathbf{H}^*(-\pi) - \mathbf{L}^*(\mathcal{A}^*\pi) \tag{28}$$

We first compute the Fenchel dual $\mathbf{H}^*(-\pi)$ and $\mathbf{L}^*(\mathcal{A}^*\pi)$ in the right-hand side of (28). For arbitrary $\pi \in \mathcal{M}(\Omega \times \Omega)$, we have

$$\begin{aligned}
&\mathbf{H}^*(-\pi) \\
&= \sup_{u \in F} \Big\{ \int (-u) d\pi - \mathbf{H}(u) \Big\} \\
&= \sup_{u \in F} \Big\{ \int (-u) d\pi |\ u(x,y) \geq -c(x,y),\ \forall (x,y) \in \Omega \times \Omega \Big\} \\
&= \sup_{u \in F} \Big\{ \int u d\pi |\ u(x,y) \leq c(x,y),\ \forall (x,y) \in \Omega \times \Omega \Big\}
\end{aligned}$$

It is easy to see that if π is a non-negative measure, then this supremum is $\int c d\pi$, otherwise it is $+\infty$. Thus

$$\mathbf{H}^*(-\pi) = \begin{cases} \int c(x,y) d\pi(x,y) & if\ \pi \in \mathcal{M}_+(\Omega \times \Omega) \\ +\infty & else \end{cases} \tag{29}$$

Similarly, we have

$$\begin{aligned}
&\mathbf{L}^*(\mathcal{A}^*\pi) \\
&= \sup_{(\boldsymbol{f},\boldsymbol{g},h) \in E} \Big\{ < (\boldsymbol{f},\boldsymbol{g},h), A^*\pi > - \mathbf{L}(\boldsymbol{f},\boldsymbol{g},h) \Big\} \\
&= \sup_{(\boldsymbol{f},\boldsymbol{g},h) \in E} \Big\{ < A(\boldsymbol{f},\boldsymbol{g},h), \pi > -(\int \boldsymbol{f} d\alpha + \int \boldsymbol{g} d\beta + hm) | \boldsymbol{f},\boldsymbol{g} \geq 0,\ h \leq 0 \Big\} \\
&= \sup_{(\boldsymbol{f},\boldsymbol{g},h) \in E} \Big\{ \int \boldsymbol{f} d(\pi^1_\# - \alpha) + \int \boldsymbol{g} d(\pi^2_\# - \beta) + h(\pi(\Omega \times \Omega) - m) | \boldsymbol{f},\boldsymbol{g} \geq 0,\ h \leq 0 \Big\}
\end{aligned}$$

If $(\alpha - \pi^1_\#)$ and $(\beta - \pi^2_\#)$ are non-negative measures, and $\pi(\Omega \times \Omega) - m \geq 0$, this supremum is 0, otherwise it is $+\infty$. Thus

$$\mathbf{L}^*(\mathcal{A}^*\pi) = \begin{cases} 0 & if (\alpha - \pi^1_\#) \in \mathcal{M}_+(\Omega),\ (\beta - \pi^2_\#) \in \mathcal{M}_+(\Omega),\ \pi(\Omega \times \Omega) \geq m \\ +\infty & else \end{cases} \tag{30}$$

In addition, the left-hand side of (28) reads

$$\begin{aligned}
&\inf_{(\boldsymbol{f},\boldsymbol{g},h) \in E} \mathbf{H}(\mathcal{A}(\mathbf{f},\mathbf{g},\mathbf{h})) + \mathbf{L}(\boldsymbol{f},\boldsymbol{g},h) \\
&= \inf_{(\boldsymbol{f},\boldsymbol{g},h) \in E} \Big\{ \int \boldsymbol{f} d\alpha + \int \boldsymbol{g} d\beta + hm |\ \boldsymbol{f},\boldsymbol{g} \geq 0,\ h \leq 0,\ \boldsymbol{f}(x) + \boldsymbol{g}(y) + h \geq -c(x,y), \forall x,y \in \Omega \Big\} \\
&= - \sup_{(\boldsymbol{f},\boldsymbol{g},h) \in E} \Big\{ \int \boldsymbol{f} d\alpha + \int \boldsymbol{g} d\beta + hm |\ \boldsymbol{f},\boldsymbol{g} \leq 0,\ h \geq 0,\ \boldsymbol{f}(x) + \boldsymbol{g}(y) + h \leq c(x,y), \forall x,y \in \Omega \Big\}
\end{aligned}$$

Finally, by inserting these terms into (28), we have

$$\sup_{\substack{(\boldsymbol{f},\boldsymbol{g},h) \in E \\ \boldsymbol{f},\boldsymbol{g} \leq 0, h \geq 0 \\ \boldsymbol{f}(x)+\boldsymbol{g}(y)+h \leq c(x,y) \forall x,y \in \Omega}} \int \boldsymbol{f} d\alpha + \int \boldsymbol{g} d\beta + hm = \inf_{\substack{\pi \in \mathcal{M}_+(\Omega \times \Omega) \\ (\alpha - \pi^1_\#) \in \mathcal{M}_+(\Omega), (\beta - \pi^2_\#) \in \mathcal{M}_+(\Omega) \\ \pi(\Omega \times \Omega) \geq m}} \int c(x,y) d\pi(x,y),$$

which proves (23).

In addition, we can also check right-hand side of (28) is finite, since we can always construct independent coupling $\widetilde{\pi} = \frac{\sqrt{m}}{\alpha(\Omega)}\alpha \otimes \frac{\sqrt{m}}{\beta(\Omega)}\beta$, such that $\widetilde{\pi} \in \mathcal{M}_+(\Omega \times \Omega)$, $\widetilde{\pi}^1_\# = \frac{m}{\alpha(\Omega)}\alpha \leq \alpha$, $\widetilde{\pi}^2_\# = \frac{m}{\beta(\Omega)}\beta \leq \beta$ and $\widetilde{\pi}(\Omega \times \Omega) = m$. Thus the Fenchel-Rockafellar duality suggests the infimum is attained. □

Proposition 4 (Dual form of $\mathcal{L}_{D,h}$)**.** *Problem 21 can be equivalently expressed as*

$$\mathcal{L}_{D,h}(\alpha,\beta) = \sup_{(\boldsymbol{f},\boldsymbol{g}) \in \mathbf{R}(h)} \int_\Omega \boldsymbol{f} d\alpha + \int_\Omega \boldsymbol{g} d\beta. \tag{31}$$

where the feasible set is

$$\mathbf{R}(h) = \left\{ (\boldsymbol{f}, \boldsymbol{g}) \in C(\Omega) \times C(\Omega) | \boldsymbol{f} \leq 0,\ \boldsymbol{g} \leq 0,\ c(x,y) - h - \boldsymbol{f}(x) - \boldsymbol{g}(y) \geq 0, \forall x, y \in \Omega \right\} \quad (32)$$

In addition, the infimum in Problem 21 is attained.

Proof. This proposition can be proved in similarly to Proposition 3, so we omit the proof here. □

By comparing Proposition 4 with Proposition 3, we can see that $\mathcal{L}_{D,h}$ and $\mathcal{L}_{M,m}$ are related by

$$\begin{aligned} \mathcal{L}_{M,m} &= \sup_{(\boldsymbol{f},\boldsymbol{g},h) \in \mathbf{R}} \int_\Omega \boldsymbol{f} d\alpha + \int_\Omega \boldsymbol{g} d\beta + mh \\ &= \sup_{h \in \mathbb{R}_+} \sup_{(\boldsymbol{f},\boldsymbol{g}) \in \mathbf{R}(h)} \int_\Omega \boldsymbol{f} d\alpha + \int_\Omega \boldsymbol{g} d\beta + mh \\ &= \sup_{h \in \mathbb{R}_+} \mathcal{L}_{D,h} + mh, \end{aligned} \quad (33)$$

which suggests that $\mathcal{L}_{M,m}$ can be regarded as $\mathcal{L}_{D,h}$ whose parameter h is chosen adaptively according to m. Thus we can derive the KR form for $\mathcal{L}_{M,m}$ by inserting (22) into (33). Formally, we have the following theorem.

Theorem 3 (KR form of $\mathcal{L}_{M,m}$)**.** *When $c(x,y) = d(x,y)$, problem 16 can be reformulated as*

$$\mathcal{L}_{M,m}(\alpha, \beta) = \sup_{\substack{\boldsymbol{f} \in Lip(\Omega), h \in \mathbf{R}_+ \\ -h \leq \boldsymbol{f} \leq 0}} \int_\Omega \boldsymbol{f} d\alpha - \int_\Omega \boldsymbol{f} d\beta + h(m - m_\beta) \quad (34)$$

or equivalently as

$$\mathcal{L}_{M,m}(\alpha, \beta) = \sup_{\boldsymbol{f} \in Lip(\Omega), \boldsymbol{f} \leq 0} \int_\Omega \boldsymbol{f} d\alpha - \int_\Omega \boldsymbol{f} d\beta - \inf(\boldsymbol{f})(m - m_\beta). \quad (35)$$

Proof. We obtain equation 34 by inserting (22) into (33) and merging two supremums. As for equation 35, note that given a fixed $\boldsymbol{f} \in C(\Omega)$ in problem 34, the optimal h is simply $-\inf(\boldsymbol{f}) < +\infty$. So we can replace h by $-\inf(\boldsymbol{f})$ in problem 34 to obtain problem 35. □

We define functionals associated to problem 34 and 35 as follows.

Definition 2 ($\mathbf{L}_{M,m}$ and $\overline{\mathbf{L}_{M,m}}$)**.** *Define functional $\mathbf{L}_{M,m}$ associated to problem 34 as*

$$\mathbf{L}_{M,m}^{\alpha,\beta}(\boldsymbol{f}, h) = \begin{cases} \int_\Omega \boldsymbol{f} d(\alpha - \beta) + h(m - m_\beta) & h \in \mathbb{R}_+\ ,\ \boldsymbol{f} \in Lip(\Omega)\ ,\ -h \leq \boldsymbol{f} \leq 0 \\ -\infty & else, \end{cases} \quad (36)$$

and functional $\overline{\mathbf{L}_{M,m}}$ associated to problem 35 as

$$\overline{\mathbf{L}_{M,m}^{\alpha,\beta}}(\boldsymbol{f}) = \begin{cases} \int_\Omega \boldsymbol{f} d(\alpha - \beta) - \inf(\boldsymbol{f})(m - m_\beta) & \boldsymbol{f} \in Lip(\Omega)\ ,\ \boldsymbol{f} \leq 0 \\ -\infty & else, \end{cases} \quad (37)$$

With these notations, problem 34 and problem 35 are simplified as

$$\mathcal{L}_{M,m}(\alpha, \beta) = \sup_{\boldsymbol{f} \in C(\Omega), h \in \mathbb{R}} \mathbf{L}_{M,m}^{\alpha,\beta}(\boldsymbol{f}, h)$$

and

$$\mathcal{L}_{M,m}(\alpha, \beta) = \sup_{\boldsymbol{f} \in C(\Omega)} \overline{\mathbf{L}_{M,m}^{\alpha,\beta}}(\boldsymbol{f})$$

respectively.

For clarity, we summarize the equivalent formulations of $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$ discussed in this work in Table. 1. We note that there are two equivalent KR forms of $\mathcal{L}_{M,m}$, and we use both of them in our theories. However, the infimum in $\overline{\mathbf{L}_{M,m}}$ is not easy to handle in practice, thus we only use $\mathbf{L}_{M,m}$ in our algorithm.

Table 1: Equivalent formulations of $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$

	$\mathcal{L}_{M,m}$	$\mathcal{L}_{D,h}$
Primal	$\inf_{\pi} \int_{\Omega\times\Omega} c(x,y)d\pi(x,y)$ $s.t.\ \pi(A\times\Omega)\le\alpha(A),\ \forall A,$ $\pi(\Omega\times A)\le\beta(A),\ \forall A,$ $\pi(\Omega\times\Omega)\ge m,$	$\inf_{\pi} \int_{\Omega\times\Omega} c(x,y)d\pi(x,y) - hm(\pi)$ $s.t.\ \pi(A\times\Omega)\le\alpha(A),\ \forall A,$ $\pi(\Omega\times A)\le\beta(A),\ \forall A,$
Dual	$\sup_{(\boldsymbol{f},\boldsymbol{g},h)} \int_{\Omega} \boldsymbol{f}d\alpha + \int_{\Omega} \boldsymbol{g}d\beta + mh$ $s.t.\ (\boldsymbol{f},\boldsymbol{g},h)\in C(\Omega)\times C(\Omega)\times\mathbb{R},$ $\boldsymbol{f}\le 0,\ \boldsymbol{g}\le 0,\ h\ge 0$ $c(x,y)-h-\boldsymbol{f}(x)-\boldsymbol{g}(y)\ge 0, \forall x,y$	$\sup_{(\boldsymbol{f},\boldsymbol{g})} \int_{\Omega} \boldsymbol{f}d\alpha + \int_{\Omega} \boldsymbol{g}d\beta$ $s.t.\ (\boldsymbol{f},\boldsymbol{g})\in C(\Omega)\times C(\Omega),$ $\boldsymbol{f}\le 0,\ \boldsymbol{g}\le 0,$ $c(x,y)-h-\boldsymbol{f}(x)-\boldsymbol{g}(y)\ge 0, \forall x,y$
KR ($c=d$)	$\sup_{\boldsymbol{f},h} \int_{\Omega} \boldsymbol{f}d(\alpha-\beta) + h(m-m_\beta)$ $s.t.\ \boldsymbol{f}\in Lip(\Omega), h\ge 0$ $-h\le\boldsymbol{f}\le 0$	$\sup_{\boldsymbol{f}} \int_{\Omega} \boldsymbol{f}d(\alpha-\beta) - hm_\beta$ $s.t.\ \boldsymbol{f}\in Lip(\Omega)$ $-h\le\boldsymbol{f}\le 0$
	$\sup_{\boldsymbol{f}} \int_{\Omega} \boldsymbol{f}d(\alpha-\beta) - \inf(\boldsymbol{f})(m-m_\beta)$ $s.t.\ \boldsymbol{f}\in Lip(\Omega)$ $\boldsymbol{f}\le 0$	

D PROPERTIES

In the previous section, we have proposed the KR forms of $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$, which are the foundations of our algorithm. To provide deeper understandings of these formulations, we investigate their properties in this section. Specifically, we first prove the solutions to the KR forms indeed exist. Then we discuss the relations amongst some equivalent formulations. We further discuss the connections amongst $\mathcal{L}_{M,m}$, $\mathcal{L}_{D,h}$ and $\mathcal{W}_1$. Finally, we present some more in-depth properties regarding the potentials of the KR forms, and the differentiability of the KR forms.

D.1 EXISTENCE OF OPTIMIZERS

As stated in Proposition 2, the maximizer of $\mathbf{L}^{\alpha,\beta}_{D,h}$ exists. In this section, we prove that the maximizer of $\mathbf{L}^{\alpha,\beta}_{M,m}$ also exists.

Lemma 1 (Continuity)**.** $\mathbf{L}^{\alpha,\beta}_{M,m}$ *is continuous on* $C(\Omega)\times\mathbb{R}$.

Proof. Let $(\boldsymbol{f}_n, h_n)\to(\boldsymbol{f},h)$ in $C(\Omega)\times\mathbb{R}$. Assume $\mathbf{L}^{\alpha,\beta}_{M,m}(\boldsymbol{f}_n,h_n)>-\infty$ when n is sufficiently large. We first check $\mathbf{L}^{\alpha,\beta}_{M,m}(\boldsymbol{f},h)>-\infty$ as follows. For arbitrary $\epsilon>0$, there exists $N>0$ such that for $n>N$, $h_n<h+\epsilon$, thus $\boldsymbol{f}_n>-h_n>-h-\epsilon$. By taking $n\to\infty$, we see for arbitrary $\epsilon>0$, $\boldsymbol{f}>-h-\epsilon$, which suggests $\boldsymbol{f}\ge -h$. In addition, it is easy to see $\boldsymbol{f}\le 0$ and $h\ge 0$. It is also easy to see $Lip(\boldsymbol{f})\le 1$ due to the closeness of $Lip(\Omega)$. Thus according to the definition 36, we claim $\mathbf{L}^{\alpha,\beta}_{M,m}(\boldsymbol{f},h)>-\infty$. Furthermore, since $-h-\epsilon<\boldsymbol{f}_n<0$, and $\boldsymbol{f}_n\to\boldsymbol{f}$, by dominated

convergence theorem, we have

$$\lim_{n\to\infty}\int_\Omega \boldsymbol{f}_n d\alpha = \int_\Omega \lim_{n\to\infty} \boldsymbol{f}_n d\alpha = \int_\Omega \boldsymbol{f} d\alpha, \tag{38}$$

$$\textit{and} \lim_{n\to\infty}\int_\Omega \boldsymbol{f}_n d\beta = \int_\Omega \lim_{n\to\infty} \boldsymbol{f}_n d\beta = \int_\Omega \boldsymbol{f} d\beta. \tag{39}$$

Note we have $h_n(m - m_\beta) \to h(m - m_\beta)$. We conclude the proof by combining these three terms and obtaining $\mathbf{L}^{\alpha,\beta}_{M,m}(\boldsymbol{f}_n, h_n) \to \mathbf{L}^{\alpha,\beta}_{M,m}(\boldsymbol{f}, h)$. □

Proposition 5 (Existence of optimizers)**.** *For $\alpha, \beta \in \mathcal{M}_+(\Omega)$ and $m > 0$, there exists $\boldsymbol{f} \in C(\Omega)$ and $h \in \mathbb{R}$ such that $\mathcal{L}_{M,m}(\alpha, \beta) = \mathbf{L}^{\alpha,\beta}_{M,m}(\boldsymbol{f}, h)$.*

Proof. If we can find a maximizing sequence $(\boldsymbol{f}_n, h_n)$ that converges to $(\boldsymbol{f}, h) \in C(\Omega) \times \mathbb{R}$, then Lemma 1 suggests that $\mathcal{L}_{M,m}(\alpha, \beta) = \sup_{\boldsymbol{f},h} \mathbf{L}^{\alpha,\beta}_{M,m}(\boldsymbol{f}, h) = \lim_{n\to\infty} \mathbf{L}^{\alpha,\beta}_{M,m}(\boldsymbol{f}_n, h_n) = \mathbf{L}^{\alpha,\beta}_{M,m}(\boldsymbol{f}, h)$, which proves this proposition. Therefore, we only need to show that it is always possible to construct such a maximizing sequence.

Let $(\boldsymbol{f}_n, h_n)$ be a maximizing sequence. We abbreviate $\max(\boldsymbol{f}_n) = \max_{x\in\Omega}(\boldsymbol{f}_n(x))$ and $\min(\boldsymbol{f}_n) = \min_{x\in\Omega}(\boldsymbol{f}_n(x))$. We first assume $\boldsymbol{f}_n$ does not have any bounded subsequence, then there exists $N > 0$, such that for all $n > N$, $\min(\boldsymbol{f}_n) < -diam(\Omega)$ (otherwise we can simply collect a subsequence of $\boldsymbol{f}_n$ bounded by $diam(\Omega)$). We can therefore construct $\widetilde{\boldsymbol{f}_n} = \boldsymbol{f}_n - (\min(\boldsymbol{f}_n) + diam(\Omega))$ and $\widetilde{h_n} = h_n + (\min(\boldsymbol{f}_n) + diam(\Omega))$. Note that $\max(\widetilde{\boldsymbol{f}_n}) \le \min(\widetilde{\boldsymbol{f}_n}) + diam(\Omega) = -diam(\Omega) + diam(\Omega) = 0$, $\widetilde{\boldsymbol{f}_n} + \widetilde{h_n} = \boldsymbol{f}_n + h_n \ge 0$, and $\widetilde{\boldsymbol{f}_n} \in Lip(\Omega)$, so $\mathbf{L}^{\alpha,\beta}_{M,m}(\widetilde{\boldsymbol{f}_n}, \widetilde{h_n}) > -\infty$, and

$$\begin{aligned}
\mathbf{L}^{\alpha,\beta}_{M,m}(\widetilde{\boldsymbol{f}_n}, \widetilde{h_n}) &= \int_\Omega \widetilde{\boldsymbol{f}_n} d(\alpha - \beta) + \widetilde{h_n}(m - m_\beta) \\
&= \int_\Omega \boldsymbol{f}_n d(\alpha - \beta) + h_n(m - m_\beta) + (\min(\boldsymbol{f}_n) - diam(\Omega))(m - m_\alpha) \\
&\ge \mathbf{L}^{\alpha,\beta}_{M,m}(\boldsymbol{f}_n, h_n),
\end{aligned}$$

which suggests that $(\widetilde{\boldsymbol{f}_n}, \widetilde{h_n})$ is a better maximizing sequence than $(\boldsymbol{f}_n, h_n)$. Note $\widetilde{\boldsymbol{f}_n}$ is uniformly bounded by $diam(\Omega)$ because $0 \ge \widetilde{\boldsymbol{f}_n} \ge \min(\widetilde{\boldsymbol{f}_n}) = -diam(\Omega)$. As a result, we can always assume $\widetilde{h_n}$ is also bounded by $diam(\Omega)$. Because otherwise we can construct $\overline{h_n} = -\min(\widetilde{\boldsymbol{f}_n}) \le diam(\Omega)$, and it is easy to show $(\widetilde{\boldsymbol{f}_n}, \overline{h_n})$ is a better maximizing sequence than $(\widetilde{\boldsymbol{f}_n}, \widetilde{h_n})$. In summary, we can always find a maximizing sequence $(\boldsymbol{f}_n, h_n)$, such that both $\boldsymbol{f}_n$ and h_n are bounded by $diam(\Omega)$.

Finally, since $\boldsymbol{f}_n$ is uniformly bounded and equicontinuous, $\boldsymbol{f}_n$ converges uniformly (up to a subsequence) to a continuous function $\boldsymbol{f}$. In addition, h_n has a convergent subsequence since it is bounded. Therefore, we can always find a maximizing sequence $(\boldsymbol{f}_n, h_n)$ that converges to some $(\boldsymbol{f}, h) \in C(\Omega) \times \mathbb{R}$, which finishes the proof. □

Remark The proof of Proposition 5 is an analogue of Proposition 2.11 in Schmitzer & Wirth (2019). The difference is that in our Proposition 5, besides $\boldsymbol{f}$, we need to handle another variable h acting as the lower bound of $\boldsymbol{f}$. In contrast, Proposition 2.11 in Schmitzer & Wirth (2019) only handles the fixed h.

D.2 Connections between the primal forms, the KR forms and the dual forms

We begin with the discussion of the connection between the dual form and the KR form for $\mathcal{L}_{M,m}$. Specifically, we have the following proposition.

Proposition 6 (KR-dual connection for $\mathcal{L}_{M,m}$)**.** *Assume $c(x, y) = d(x, y)$.*

1. *If $(\boldsymbol{f}, h)$ is a maximizer of $\mathbf{L}^{\alpha,\beta}_{M,m}$, then $(\boldsymbol{f}, -\boldsymbol{f} - h, h)$ and $(-\boldsymbol{f} - h, \boldsymbol{f}, h)$ are solutions to the corresponding dual form 23.*

2. *If $(\boldsymbol{f}, \boldsymbol{g}, h)$ is a solution to the dual form 23, then $(\boldsymbol{g}^{\overline{c(h)}}, h)$ and $(-h - \boldsymbol{f}^{\overline{c(h)}}, h)$ are the maximizers of the corresponding $\mathbf{L}_{M,m}^{\alpha,\beta}$, where $\boldsymbol{g}^{\overline{c(h)}}$ is the $\overline{c(h)}$-transformed $\boldsymbol{g}$, and is defined as*

$$\boldsymbol{g}^{\overline{c(h)}}(y) = \min\Big\{\inf_x \Big(c(x,y) - h - \boldsymbol{g}(x)\Big), 0\Big\}. \tag{40}$$

The proof of the second part of Proposition 6 requires a few lemmas regarding the $\overline{c(h)}$-transform.

Lemma 2. *Given functions $\boldsymbol{f}$ and $\boldsymbol{g}$ defined on Ω, we always have $|\inf \boldsymbol{f} - \inf \boldsymbol{g}| \leq \sup |\boldsymbol{f} - \boldsymbol{g}|$.*

Lemma 3. *When $c(x,y) = d(x,y)$, given a function $\boldsymbol{f} \leq 0$ and $h \in \mathbb{R}_+$, $\boldsymbol{f}^{\overline{c(h)}}$ has the following two properties, 1) $-h \leq \boldsymbol{f}^{\overline{c(h)}} \leq 0$. 2) $\boldsymbol{f}^{\overline{c(h)}} \in Lip(\Omega)$*

Proof. 1) On one hand, it is easy to see $\boldsymbol{f}^{\overline{c(h)}} \leq 0$ by definition. On the other hand, by inserting $d(x,y) \geq 0$ and $\boldsymbol{f} \leq 0$ into (40), we have

$$\begin{aligned}
\boldsymbol{f}^{\overline{c(h)}} &= \min\Big\{\inf_x \Big(d(x,y) - h - \boldsymbol{f}(x)\Big), 0\Big\} \\
&\geq \min\Big\{\inf_x \Big(0 - h - 0\Big), 0\Big\} \\
&= -h.
\end{aligned}$$

2) Considering the function $\overline{\boldsymbol{f}} = \inf_x \Big(d(x,y) - h - \boldsymbol{f}(x)\Big)$, we always have $\overline{\boldsymbol{f}} \in Lip(\Omega)$, because for arbitrary $y_1, y_2 \in \Omega$,

$$\begin{aligned}
|\overline{\boldsymbol{f}}(y_1) - \overline{\boldsymbol{f}}(y_2)| &= |\inf_x \Big(d(x,y_1) - h - \boldsymbol{f}(x)\Big) - \inf_x \Big(d(x,y_2) - h - \boldsymbol{f}(x)\Big)| \\
&\leq \sup_x |d(x,y_1) - d(x,y_2)| && \text{(by Lemma 2)} \\
&\leq d(y_1,y_2), && \text{(triangular inequality)}
\end{aligned}$$

Furthermore, it is straightforward to check that

$$\begin{aligned}
|\boldsymbol{f}^{\overline{c(h)}}(y_1) - \boldsymbol{f}^{\overline{c(h)}}(y_2)| = |\min\Big\{\overline{\boldsymbol{f}}(y_1), 0\Big\} - \min\Big\{\overline{\boldsymbol{f}}(y_2), 0\Big\}| &\leq |\overline{\boldsymbol{f}}(y_1) - \overline{\boldsymbol{f}}(y_2)| \\
&\leq d(y_1,y_2),
\end{aligned}$$

which suggests $\boldsymbol{f}^{\overline{c(h)}} \in Lip(\Omega)$. □

Lemma 4. *Assume $c(x,y) = d(x,y)$. Given a function $\boldsymbol{f}$ satisfying the following two properties 1) $-h \leq \boldsymbol{f} \leq 0$, 2) $\boldsymbol{f} \in Lip(\Omega)$, the following two statements hold. 1) $\boldsymbol{f}^{\overline{c(h)}}$ can be explicitly written as $\boldsymbol{f}^{\overline{c(h)}} = -h - \boldsymbol{f}$. 2) There exists a function $\boldsymbol{g}$, such that $\boldsymbol{f} = \boldsymbol{g}^{\overline{c(h)}}$.*

Proof. 1) Since $\boldsymbol{f} \in Lip(\Omega)$, for arbitrary $x, y \in \Omega$, $\boldsymbol{f}(y) - d(x,y) \leq \boldsymbol{f}(x) \leq \boldsymbol{f}(y) + d(x,y)$. Therefore,

$$\begin{aligned}
\boldsymbol{f}^{\overline{c(h)}}(y) &= \min\Big\{\inf_x \Big(d(x,y) - h - \boldsymbol{f}(x)\Big), 0\Big\} \\
&\geq \min\Big\{\inf_x \Big(d(x,y) - h - \boldsymbol{f}(y) - d(x,y)\Big), 0\Big\} \\
&= \min\Big\{-\boldsymbol{f}(y) - h, 0\Big\} \\
&= -\boldsymbol{f}(y) - h.
\end{aligned}$$

Not the last equation holds because $-\boldsymbol{f} - h \leq 0$. Similarly, we have

$$\begin{aligned}
\boldsymbol{f}^{\overline{c(h)}}(y) &= \min\Big\{\inf_x \Big(d(x,y) - h - \boldsymbol{f}(x)\Big), 0\Big\} \\
&\leq \min\Big\{\inf_x \Big(d(x,y) - h - \boldsymbol{f}(y) + d(x,y)\Big), 0\Big\} \\
&= \min\Big\{-\boldsymbol{f}(y) - h, 0\Big\} \\
&= -\boldsymbol{f}(y) - h.
\end{aligned}$$

Therefore, we have $\boldsymbol{f}^{\overline{c(h)}} = -h - \boldsymbol{f}$

2) We construct $\boldsymbol{g} = -\boldsymbol{f} - h$. It is easy to check $-h \leq \boldsymbol{g} \leq 0$ and $\boldsymbol{g} \in Lip(\Omega)$ According to the first part of this proposition, we claim $\boldsymbol{g}^{\overline{c(h)}} = -h - \boldsymbol{g} = \boldsymbol{f}$, which concludes the proof. □

Now we prove Proposition 6.

Proof. (1) We only prove $(\boldsymbol{f}, -\boldsymbol{f} - h, h)$ is a solution to the dual form 23, and the proof of $(-\boldsymbol{f} - h, \boldsymbol{f}, h)$ is similar.

First, we check $\boldsymbol{f} \leq 0$, $-\boldsymbol{f} - h \leq 0$ and $h \geq 0$. In addition, we have

$$d(x, y) - h - \boldsymbol{f}(x) - (-\boldsymbol{f}(y) - h)) = d(x, y) - (\boldsymbol{f}(x) - \boldsymbol{f}(y)) \geq 0$$

where the final inequality holds due to $\boldsymbol{f} \in Lip(\Omega)$. Thus $(\boldsymbol{f}, -\boldsymbol{f} - h, h) \in \mathbf{R}$. Finally, by re-arranging the terms, we have

$$\int_\Omega \boldsymbol{f} d\alpha + \int_\Omega (-\boldsymbol{f} - h) d\beta + mh = \mathbf{L}^{\alpha,\beta}_{M,m}(\boldsymbol{f}, h).$$

By assumption, $(\boldsymbol{f}, h)$ is the maximizer of $\mathbf{L}^{\alpha,\beta}_{M,m}$, thus $\mathbf{L}^{\alpha,\beta}_{M,m}(\boldsymbol{f}, h) = \mathcal{L}_{M,m}(\alpha, \beta)$, which completes our proof.

(2) We only prove $(-h - \boldsymbol{f}^{\overline{c(h)}}, h)$ is a maximizer of $\mathbf{L}^{\alpha,\beta}_{M,m}$, and the proof of $(\boldsymbol{g}^{\overline{c(h)}}, h)$ is similar.

For a solution $(\boldsymbol{f}, \boldsymbol{g}, h) \in \mathbf{R}$ to the dual form 23, it is easy to check $(\boldsymbol{f}, \boldsymbol{f}^{\overline{c(h)}}, h) \in \mathbf{R}$. Because we have $\boldsymbol{f}^{\overline{c(h)}} \in Lip(\Omega) \subseteq C(\Omega)$, $\boldsymbol{f}^{\overline{c(h)}} \leq 0$ according to Lemma 3, and $d(x, y) - h - \boldsymbol{f} - \boldsymbol{f}^{\overline{c(h)}} \geq 0$ by definition. In addtion, for arbitrary $x, y \in \Omega$, we have $\boldsymbol{g}(y) \leq d(x, y) - h - \boldsymbol{f}(x)$, thus $\boldsymbol{g}(y) \leq \inf_x\{d(x, y) - h - \boldsymbol{f}(x)\}$. Combining this and the fact $\boldsymbol{g} \leq 0$, we obtain $\boldsymbol{g}(y) \leq \boldsymbol{f}^{\overline{c(h)}}$. Since β is a non-negative measure, we have

$$\int_\Omega \boldsymbol{f} d\alpha + \int_\Omega \boldsymbol{g} d\beta + mh \leq \int_\Omega \boldsymbol{f} d\alpha + \int_\Omega \boldsymbol{f}^{\overline{c(h)}} d\beta + mh,$$

which suggests $(\boldsymbol{f}, \boldsymbol{f}^{\overline{c(h)}}, h)$ is also a solution to the dual form 23. By repeating this argument, we claim $((\boldsymbol{f}^{\overline{c(h)}})^{\overline{c(h)}}, \boldsymbol{f}^{\overline{c(h)}}, h)$ is also a solution. According to Lemma 3, $-h \leq \boldsymbol{f}^{\overline{c(h)}} \leq 0$ and $\boldsymbol{f}^{\overline{c(h)}} \in Lip(\Omega)$, therefore Lemma 4 suggests $(\boldsymbol{f}^{\overline{c(h)}})^{\overline{c(h)}} = -\boldsymbol{f}^{\overline{c(h)}} - h$, *i.e.*, $(-\boldsymbol{f}^{\overline{c(h)}} - h, \boldsymbol{f}^{\overline{c(h)}}, h) = ((\boldsymbol{f}^{\overline{c(h)}})^{\overline{c(h)}}, \boldsymbol{f}^{\overline{c(h)}}, h)$ is a solution, In other words,

$$\mathcal{L}_{M,m}(\alpha, \beta) = \int_\Omega (-\boldsymbol{f}^{\overline{c(h)}} - h) \boldsymbol{f} d\alpha + \int_\Omega \boldsymbol{f}^{\overline{c(h)}} d\beta + mh.$$

Furthermore, since $(-\boldsymbol{f}^{\overline{c(h)}} - h) \in Lip(\Omega)$ and $-h \leq (-\boldsymbol{f}^{\overline{c(h)}} - h) \leq 0$, the right-hand side of this equations equals $\mathbf{L}^{\alpha,\beta}_{M,m}(-\boldsymbol{f}^{\overline{c(h)}} - h, h)$, which finishes our proof. □

Remark It is worth noticing that Proposition 6 actually provide another way to construct the respective KR forms of $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$ without the aid of KR metric. Actually, the proof of Proposition 6 is an analogue of the classic Kantorovich-Rubinstein duality, where the role played by $\overline{c(h)}$-transform in our theory is similar to the role played by c-transform in classic OT literatures, *e.g.*, Villani (2009).

We note that from the proof of Proposition 6, we can see that the solution to the dual form can be always written as $(\boldsymbol{f}, \boldsymbol{f}^{\overline{c(h)}}, h)$, where $(\boldsymbol{f}, h)$ are the only two unknown variables.

We now present the relations between the dual form and the primal form.

Proposition 7 (Dual-primal connection for $\mathcal{L}_{M,m}$)**.** *Let $(\boldsymbol{f}, \boldsymbol{f}^{\overline{c(h)}}, h)$ be a solution to the dual form 31, and π be the solution to the corresponding primal form 16. We have*

$$\int_{\Omega\times\Omega} (c(x, y) - \boldsymbol{f}(x) - \boldsymbol{f}^{\overline{c(h)}}(y) - h) d\pi(x, y) = 0 \tag{41}$$

Proof. According to the equivalence of the dual and primal form, we have

$$\int_{\Omega\times\Omega} c(x,y)d\pi(x,y) = \int_{\Omega} \boldsymbol{f} d\alpha + \int_{\Omega} \boldsymbol{f}^{\overline{c(h)}} d\beta + mh.$$

Since $\boldsymbol{f} \leq 0$ and $\pi^1_{\#} \leq \alpha$, we have $\int_\Omega \boldsymbol{f} d\alpha \leq \int_\Omega \boldsymbol{f} d\pi^1_{\#}$. Similarly, we have $\int_\Omega \boldsymbol{f}^{\overline{c(h)}} d\beta \leq \int_\Omega \boldsymbol{f}^{\overline{c(h)}} d\pi^2_{\#}$. In addition, we have $\pi(\Omega\times\Omega) \geq m$ and $h \geq 0$, so we have $hm \leq \int_{\Omega\times\Omega} h d\pi$. By inserting these inequalities into the above equation, we have

$$\int_{\Omega\times\Omega} c(x,y)d\pi(x,y) \leq \int_{\Omega} \boldsymbol{f}(x) d\pi^1_{\#}(x) + \int_{\Omega} \boldsymbol{f}^{\overline{c(h)}}(y) d\pi^2_{\#}(y) + \int_{\Omega\times\Omega} h d\pi.$$

By re-arranging this inequality, we have

$$\int_{\Omega\times\Omega} (c(x,y) - \boldsymbol{f}(x) - \boldsymbol{f}^{\overline{c(h)}}(y) - h) d\pi(x,y) \leq 0.$$

In addition, since $(\boldsymbol{f}, \boldsymbol{f}^{\overline{c(h)}}, h) \in \mathbf{R}$, we have $c(x,y) - \boldsymbol{f}(x) - \boldsymbol{f}^{\overline{c(h)}}(y) - h \geq 0$, thus

$$\int_{\Omega\times\Omega} (c(x,y) - \boldsymbol{f}(x) - \boldsymbol{f}^{\overline{c(h)}}(y) - h) d\pi(x,y) \geq 0,$$

which suggests the equality is obtained, and finishes the proof. □

As a direct consequence, we obtain the connection between the primal and the KR form for $\mathcal{L}_{M,m}$.

Corollary 1 (Primal-KR relation for $\mathcal{L}_{M,m}$)**.** *Assume $c(x,y) = d(x,y)$. Let $(\boldsymbol{f}, h)$ be a maximizer of $\mathbf{L}_{M,m}$, and π be the solution to the corresponding primal form 16. We have*

$$\int_{\Omega\times\Omega} (d(x,y) - \boldsymbol{f}(x) + \boldsymbol{f}(y)) d\pi(x,y) = 0. \tag{42}$$

Proof. According to Lemma 4, we have $\boldsymbol{f}^{\overline{c(h)}} = -h - \boldsymbol{f}$. Thus this corollary is a direct result of Proposition 7. □

We now present the results for $\mathcal{L}_{D,h}$. We omit the proofs for these results since they can be proved in a similar way as their counterparts for $\mathcal{L}_{M,m}$.

Proposition 8 (KR-dual connection of $\mathcal{L}_{D,h}$)**.** *Assume $c(x,y) = d(x,y)$.*

1. *If $\boldsymbol{f}$ is a maximizer of $\mathbf{L}^{\alpha,\beta}_{D,h}$, then $(\boldsymbol{f}, -\boldsymbol{f} - h, h)$ and $(-\boldsymbol{f} - h, \boldsymbol{f}, h)$ are solutions to the corresponding dual form 31.*

2. *If $(\boldsymbol{f}, \boldsymbol{g}, h)$ is a solution to the dual form 31, then $\boldsymbol{g}^{\overline{c(h)}}$ and $-h - \boldsymbol{f}^{\overline{c(h)}}$ are maximizers of the corresponding $\mathbf{L}^{\alpha,\beta}_{D,h}$, where $(\cdot)^{\overline{c(h)}}$ is the $\overline{c(h)}$-transform defined in (40),*

Proposition 9 (Dual-primal connection for $\mathcal{L}_{D,h}$)**.** *Let $(\boldsymbol{f}, \boldsymbol{f}^{\overline{c(h)}})$ be a solution to the dual form 31, and π be the solution to the corresponding primal form 16. We have*

$$\int_{\Omega\times\Omega} (c(x,y) - \boldsymbol{f}(x) - \boldsymbol{f}^{\overline{c(h)}}(y) - h) d\pi(x,y) = 0 \tag{43}$$

Proposition 10 (Primal-KR relation for $\mathcal{L}_{D,h}$)**.** *Assume $c(x,y) = d(x,y)$. Let $\boldsymbol{f}$ be a maximizer of $\mathbf{L}_{D,h}$, and π be the solution to the corresponding primal form 16. We have*

$$\int_{\Omega\times\Omega} (d(x,y) - \boldsymbol{f}(x) + \boldsymbol{f}(y)) d\pi(x,y) = 0. \tag{44}$$

D.3 CONNECTIONS BETWEEN $\mathcal{L}_M$, $\mathcal{L}_D$ AND $\mathcal{W}_1$

We first explain the relations between the KR forms of $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$, and then we discuss their relations to $\mathcal{W}_1$.

As can be seen in (33), the maximizers of $\mathbf{L}_{D,h}^{\alpha,\beta}$ recovers to that of $\overline{\mathbf{L}_{M,m}^{\alpha,\beta}}$ when h is chosen properly. We re-state this observation formally in the following proposition.

Proposition 11 (Relations between $\mathcal{L}_M$ and $\mathcal{L}_D$)**.** *Let $\boldsymbol{f}$ be a maximizer of $\overline{\mathbf{L}_{M,m}^{\alpha,\beta}}$. For the fixed $h^* = -\inf(\boldsymbol{f})$, $\boldsymbol{f}$ is also a maximizer of $\mathbf{L}_{D,h^*}^{\alpha,\beta}$.*

Proof. First, it is easy to check that by definition, for arbitrary $\boldsymbol{f} \in C(\Omega)$, $\mathbf{L}_{D,-\inf(\boldsymbol{f})}^{\alpha,\beta}(\boldsymbol{f}) - \inf(\boldsymbol{f})m = \overline{\mathbf{L}_{M,m}^{\alpha,\beta}}(\boldsymbol{f})$. Then we proceed by contradiction. If $\boldsymbol{f}$ is not a maximizer of $\mathbf{L}_{D,h^*}^{\alpha,\beta}$, then there exists $\overline{\boldsymbol{f}} \in Lip(\Omega)$ such that $\mathbf{L}_{D,-\inf(\boldsymbol{f})}^{\alpha,\beta}(\overline{\boldsymbol{f}}) > \mathbf{L}_{D,-\inf(\boldsymbol{f})}^{\alpha,\beta}(\boldsymbol{f}) > -\infty$, thus $\inf(\overline{\boldsymbol{f}}) \geq \inf(\boldsymbol{f})$, and we have

$$\begin{aligned}
\overline{\mathbf{L}_{M,m}^{\alpha,\beta}}(\overline{\boldsymbol{f}}) &= \int_\Omega \overline{\boldsymbol{f}} d\alpha - \int_\Omega \overline{\boldsymbol{f}} d\beta + \inf(\overline{\boldsymbol{f}})(m_\beta - m) \\
&\geq \int_\Omega \overline{\boldsymbol{f}} d\alpha - \int_\Omega \overline{\boldsymbol{f}} d\beta + \inf(\boldsymbol{f})(m_\beta - m) \\
&= \mathbf{L}_{D,-\inf(\boldsymbol{f})}^{\alpha,\beta}(\overline{\boldsymbol{f}}) - \inf(\boldsymbol{f})m \\
&> \mathbf{L}_{D,-\inf(\boldsymbol{f})}^{\alpha,\beta}(\boldsymbol{f}) - \inf(\boldsymbol{f})m \\
&= \overline{\mathbf{L}_{M,m}^{\alpha,\beta}}(\boldsymbol{f}),
\end{aligned}$$

which contradicts the assumption that $\boldsymbol{f}$ is a maximizer of $\overline{\mathbf{L}_{M,m}^{\alpha,\beta}}$. □

Now we turn to the special cases of $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$. To begin with, we define two KR forms as follows.

Definition 3 ($\mathbf{L}_P^{\alpha,\beta}$)**.** *Define the functional $\mathbf{L}_P^{\alpha,\beta}$ as*

$$\mathbf{L}_P^{\alpha,\beta}(\boldsymbol{f}) = \begin{cases} \int_\Omega \boldsymbol{f} d\alpha - \int_\Omega \boldsymbol{f} d\beta & \boldsymbol{f} \in Lip(\Omega)\ ,\ \boldsymbol{f} \leq 0,\ m_\alpha \geq m_\beta \\ -\infty & else, \end{cases} \tag{45}$$

and define $\mathcal{L}_P(\alpha,\beta) = \sup_{\boldsymbol{f}\in C(\Omega)} \mathbf{L}_P^{\alpha,\beta}(\boldsymbol{f})$.

Definition 4 ($\mathbf{W}_1^{\alpha,\beta}$)**.** *Define the functional associated to Wasserstein-1 metric $\mathcal{W}_1$ as*

$$\mathbf{L}_W^{\alpha,\beta}(\boldsymbol{f}) = \begin{cases} \int_\Omega \boldsymbol{f} d\alpha - \int_\Omega \boldsymbol{f} d\beta & \boldsymbol{f} \in Lip(\Omega)\ m_\alpha \geq m_\beta \\ -\infty & else, \end{cases} \tag{46}$$

thus $\mathcal{W}_1$ can be expressed as $\mathcal{W}_1(\alpha,\beta) = \sup_{\boldsymbol{f}\in C(\Omega)} \mathbf{L}_W^{\alpha,\beta}(\boldsymbol{f})$. We note the maximizer of $\mathbf{L}_W^{\alpha,\beta}$ is known to exist due to Kantorovich–Rubinstein duality.

The following proposition describes the relations between $\mathcal{L}_{M,m}$, $\mathcal{L}_{D,h}$, $\mathcal{L}_P$ and $\mathcal{W}_1$.

Proposition 12 (Special cases of $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$)**.** *(1) When $m_\alpha \geq m_\beta = m$, $\mathcal{L}_{M,m}$ can be equivalently expressed as $\mathcal{L}_P$.*

(2) When $m_\alpha \geq m_\beta$ and $h \geq diam(\Omega)$, $\mathcal{L}_{D,h}$ can be equivalently expressed as $\mathcal{L}_P - hm_\beta$.

(3) When $m_\alpha = m_\beta$, $\mathcal{L}_P$ becomes equivalent to $\mathcal{W}_1$ defined in (17).

Proof. (1) It is trivial to check when $m_\alpha \geq m_\beta = m$, $\mathbf{L}_{M,m}$ becomes $\mathbf{L}_P$.

(2) Let $\boldsymbol{f}$ be a maximizer of $\mathbf{L}_{D,h}^{\alpha,\beta}$. We only need to prove that $\boldsymbol{f}$ is also a maximizer of $\mathbf{L}_P^{\alpha,\beta}$, because this implies $\mathcal{L}_{D,h}(\alpha,\beta) = \mathbf{L}_{D,h}^{\alpha,\beta}(\boldsymbol{f}) = \mathbf{L}_P^{\alpha,\beta}(\boldsymbol{f}) - hm_\beta = \mathcal{L}_P(\alpha,\beta) - hm_\beta$. We proceed by contradiction.

Assume there exists $\overline{\boldsymbol{f}}$ such that $\mathbf{L}_P^{\alpha,\beta}(\overline{\boldsymbol{f}}) > \mathbf{L}_P^{\alpha,\beta}(\boldsymbol{f})$. We construct $\widetilde{\boldsymbol{f}} = \overline{\boldsymbol{f}} - \sup(\overline{\boldsymbol{f}})$. Since $m_\alpha \geq m_\beta$, we have $\mathbf{L}_P^{\alpha,\beta}(\widetilde{\boldsymbol{f}}) \geq \mathbf{L}_P^{\alpha,\beta}(\overline{\boldsymbol{f}}) > \mathbf{L}_P^{\alpha,\beta}(\boldsymbol{f})$. In addition, since $\widetilde{\boldsymbol{f}} \in Lip(\Omega)$, we have $\inf(\widetilde{\boldsymbol{f}}) \geq \sup(\widetilde{\boldsymbol{f}}) - diam(\Omega) = -diam(\Omega) \geq -h$, thus $\mathbf{L}_{D,h}^{\alpha,\beta}(\widetilde{\boldsymbol{f}}) > -\infty$ and $\mathbf{L}_{D,h}^{\alpha,\beta}(\widetilde{\boldsymbol{f}}) = \mathbf{L}_P^{\alpha,\beta}(\widetilde{\boldsymbol{f}}) - hm_\beta > \mathbf{L}_P^{\alpha,\beta}(\boldsymbol{f}) - hm_\beta = \mathbf{L}_{D,h}^{\alpha,\beta}(\boldsymbol{f})$, which contradicts the assumption that $\boldsymbol{f}$ is a maximizer of $\mathbf{L}_{D,h}^{\alpha,\beta}$.

(3) Let $\boldsymbol{f}$ be a maximizer of $\mathbf{L}_W^{\alpha,\beta}$. We construct $\widetilde{\boldsymbol{f}} = \boldsymbol{f} - \sup(\boldsymbol{f})$. It is easy to see $\mathbf{L}_W^{\alpha,\beta}(\boldsymbol{f}) = \mathbf{L}_W^{\alpha,\beta}(\widetilde{\boldsymbol{f}})$, thus $\widetilde{\boldsymbol{f}}$ is also a maximizer of $\mathbf{L}_W^{\alpha,\beta}$. We claim $\widetilde{\boldsymbol{f}}$ is a maximizer of $\mathbf{L}_P^{\alpha,\beta}$, because otherwise there exists $\overline{\boldsymbol{f}}$ such that $\mathbf{L}_P^{\alpha,\beta}(\overline{\boldsymbol{f}}) > \mathbf{L}_P^{\alpha,\beta}(\widetilde{\boldsymbol{f}})$, thus $\mathbf{L}_W^{\alpha,\beta}(\overline{\boldsymbol{f}}) = \mathbf{L}_P^{\alpha,\beta}(\overline{\boldsymbol{f}}) > \mathbf{L}_P^{\alpha,\beta}(\widetilde{\boldsymbol{f}}) = \mathbf{L}_W^{\alpha,\beta}(\widetilde{\boldsymbol{f}})$, which contradicts the assumption that $\widetilde{\boldsymbol{f}}$ is a maximizer of $\mathbf{L}_W^{\alpha,\beta}$. As a result, we have $\mathcal{W}_1(\alpha,\beta) = \mathbf{L}_W^{\alpha,\beta}(\widetilde{\boldsymbol{f}}) = \mathbf{L}_P^{\alpha,\beta}(\widetilde{\boldsymbol{f}}) = \mathcal{L}_P(\alpha,\beta)$. □

Remark The significance of Proposition 12 is in two folds. On the one hand, it provides a new KR form $\mathcal{L}_P$ for the "semi-complete" Wasserstein problem, where all mass of β is transported to α. This problem has many practical applications, such as the template matching problem, where all data points in an incomplete "data distribution" is matched to a complete "model distribution". Thus $\mathcal{L}_P$ may be interesting on its own in those applications. We note Fig. 9(c) is an example of this case, because all mass in β is transferred. On the other hand, Proposition 12 clearly suggests that both $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$ generalize $\mathcal{W}_1$. We state it formally in the following corollary.

Corollary 2. *(1) When $m_\alpha = m_\beta = m$, $\mathcal{L}_{M,m}$ is equivalent to $\mathcal{W}_1$.*

(2) When $m_\alpha = m_\beta$ and $h \geq diam(\Omega)$, $\mathcal{L}_{D,h}$ is equivalent to $\mathcal{W}_1 - hm_\beta$.

D.4 Properties of the potentials

This subsection characterizes the maximizers of $\overline{\mathbf{L}_{M,m}}$ and $\mathbf{L}_{D,h}$ in details. We always assume $c = d$ in this subsection.

D.4.1 Mass-type PW discrepancy

First of all, it is easy to see that the maximizers of $\overline{\mathbf{L}_{M,m}}$ generally lie on the boundary of the feasible set. Formally, we have the following proposition.

Proposition 13. *Let $\boldsymbol{f}$ denote a maximizer of $\overline{\mathbf{L}_{M,m}}$.*

(1) If $m_\alpha > m$, $\sup(\boldsymbol{f}) = 0$.

(2) If $\mathcal{L}_{M,m}(\alpha,\beta) > 0$, $Lip(\boldsymbol{f}) = 1$.

Proof. The proof is straightforward. (1) If $\sup(\boldsymbol{f}) < 0$, we construct $\tilde{\boldsymbol{f}} = \boldsymbol{f} - \sup(\boldsymbol{f})$. It is easy to check $\overline{\mathbf{L}_{M,m}}(\tilde{\boldsymbol{f}}) > -\infty$ and $\overline{\mathbf{L}_{M,m}}(\tilde{\boldsymbol{f}}) - \overline{\mathbf{L}_{M,m}}(\boldsymbol{f}) = -\sup(\boldsymbol{f})(m_\alpha - m) > 0$, which contradicts the optimality of $\boldsymbol{f}$. (2) Similarly, if $Lip(\boldsymbol{f}) < 1$, we can construct $\tilde{\boldsymbol{f}} = \boldsymbol{f}/Lip(\boldsymbol{f})$. It is easy to check $\overline{\mathbf{L}_{M,m}}(\tilde{\boldsymbol{f}}) > -\infty$ and $\overline{\mathbf{L}_{M,m}}(\tilde{\boldsymbol{f}}) = \overline{\mathbf{L}_{M,m}}(\boldsymbol{f})/Lip(\boldsymbol{f}) > \overline{\mathbf{L}_{M,m}}(\boldsymbol{f})$, which contradicts the optimality of $\boldsymbol{f}$. □

We then discuss the symmetry of $\overline{\mathbf{L}_{M,m}}$. Since the functional $\overline{\mathbf{L}_{M,m}}$ is not symmetric w.r.t. α and β, *i.e.*, $\overline{\mathbf{L}_{M,m}^{\alpha,\beta}}(\boldsymbol{f}) \neq \overline{\mathbf{L}_{M,m}^{\beta,\alpha}}(\boldsymbol{f})$, it is natural to ask what is the relation between the maximizer of $\overline{\mathbf{L}_{M,m}^{\alpha,\beta}}$ and that of $\overline{\mathbf{L}_{M,m}^{\beta,\alpha}}$ and whether we have $\mathcal{L}_{M,m}(\alpha,\beta) = \mathcal{L}_{M,m}(\beta,\alpha)$. These questions are answered in the following proposition.

Proposition 14 (Symmetry)**.** *Let $\boldsymbol{f}$ denote a maximizer of $\overline{\mathbf{L}_{M,m}^{\alpha,\beta}}$. $\tilde{\boldsymbol{f}} = -\boldsymbol{f} + \inf(\boldsymbol{f})$ is a maximizer of $\overline{\mathbf{L}_{M,m}^{\beta,\alpha}}$, In addition, we have $\mathcal{L}_{M,m}(\alpha,\beta) = \mathcal{L}_{M,m}(\beta,\alpha)$.*

Proof. For the easy of notation, we abbreviate $\overline{\mathbf{L}^{\alpha,\beta}} = \overline{\mathbf{L}_{M,m}^{\alpha,\beta}}$. Consider $\boldsymbol{f}' = \boldsymbol{f} - \sup(\boldsymbol{f})$. It is easy to check $\overline{\mathbf{L}^{\alpha,\beta}}(\boldsymbol{f}') \geq \overline{\mathbf{L}^{\alpha,\beta}}(\boldsymbol{f})$, thus $\boldsymbol{f}'$ is also a maximizer of $\overline{\mathbf{L}^{\alpha,\beta}}$. Now we prove $\overline{\mathbf{L}^{\beta,\alpha}}(\tilde{\boldsymbol{f}}) =$

$\overline{\mathbf{L}^{\alpha,\beta}}(\boldsymbol{f}')$ as follows. First, we have $\overline{\mathbf{L}^{\beta,\alpha}}(\tilde{\boldsymbol{f}}) > -\infty$, because $\tilde{\boldsymbol{f}} \leq -\inf(\boldsymbol{f}) + \inf(\boldsymbol{f}) = 0$, and $Lip(\tilde{\boldsymbol{f}}) = Lip(\boldsymbol{f}) \leq 1$. Then, since $\inf(\tilde{\boldsymbol{f}}) = -sup(\boldsymbol{f}) + \inf(\boldsymbol{f}) = \inf(\boldsymbol{f}')$, we have

$$\begin{aligned}
\overline{\mathbf{L}^{\beta,\alpha}}(\tilde{\boldsymbol{f}}) &= \int_\Omega \tilde{\boldsymbol{f}} d\beta - \int_\Omega \tilde{\boldsymbol{f}} d\alpha - \inf(\tilde{\boldsymbol{f}})(m - m_\alpha) \\
&= \int_\Omega (-\boldsymbol{f} + \inf(\boldsymbol{f})) d\beta - \int_\Omega (-\boldsymbol{f} + \inf(\boldsymbol{f})) d\alpha - (\inf(\boldsymbol{f}) - \sup(\boldsymbol{f}))(m - m_\alpha) \\
&= \int_\Omega (\boldsymbol{f} - \sup(\boldsymbol{f})) d\alpha - \int_\Omega (\boldsymbol{f} - \sup(\boldsymbol{f})) d\beta - (\inf(\boldsymbol{f}) - \sup(\boldsymbol{f}))(m - m_\beta) \\
&= \int_\Omega \boldsymbol{f}' d\alpha - \int_\Omega \boldsymbol{f}' d\beta - \inf(\boldsymbol{f}')(m - m_\beta) \\
&= \overline{\mathbf{L}^{\alpha,\beta}}(\boldsymbol{f}').
\end{aligned}$$

Finally, we complete the proof by contradiction. If $\tilde{\boldsymbol{f}}$ is not a maximizer of $\overline{\mathbf{L}^{\beta,\alpha}}$, then there exists $\boldsymbol{f}^*$ such that $\overline{\mathbf{L}^{\beta,\alpha}}(\boldsymbol{f}^*) > \overline{\mathbf{L}^{\beta,\alpha}}(\tilde{\boldsymbol{f}})$. By the above proof, we can construct $\overline{\boldsymbol{f}} = -\boldsymbol{f}^* + \inf(\boldsymbol{f}^*)$ such that $\overline{\mathbf{L}^{\beta,\alpha}}(\boldsymbol{f}^*) = \overline{\mathbf{L}^{\alpha,\beta}}(\overline{\boldsymbol{f}})$, thus $\overline{\mathbf{L}^{\alpha,\beta}}(\overline{\boldsymbol{f}}) = \overline{\mathbf{L}^{\beta,\alpha}}(\boldsymbol{f}^*) > \overline{\mathbf{L}^{\beta,\alpha}}(\tilde{\boldsymbol{f}}) = \overline{\mathbf{L}^{\alpha,\beta}}(\boldsymbol{f}')$, which contradicts the optimality of $\boldsymbol{f}'$. In addition, since we have proved $\tilde{\boldsymbol{f}}$ is a maximizer of $\overline{\mathbf{L}^{\beta,\alpha}}$, we have $\mathcal{L}_{M,m}(\beta,\alpha) = \overline{\mathbf{L}^{\beta,\alpha}}(\tilde{\boldsymbol{f}}) = \overline{\mathbf{L}^{\alpha,\beta}}(\boldsymbol{f}') = \mathcal{L}_{M,m}(\alpha,\beta)$. □

Finally, we characterize the "inactive" mass of $\mathcal{L}_{M,m}$. One important property of $\mathcal{L}_{M,m}$ is that it has some "inactive" mass that is not transported. A direct observation is that if we throw away some inactive mass, the solution to the problem will not change. This observation is formally stated as follows.

Proposition 15 (Inactive mass)**.** *Let π be the solution to the primal form of $\mathcal{L}_{M,m}(\alpha,\beta)$. Define α' and β' be measures satisfying*

$$\pi^1_\#(A) \leq \alpha'(A) \leq \alpha(A) \quad \textit{and} \quad \pi^2_\#(A) \leq \beta'(A) \leq \beta(A) \tag{47}$$

for an arbitrary measurable set A.

(1) π is also the solution to the primal form of $\mathcal{L}_{M,m}(\alpha',\beta')$, thus $\mathcal{L}_{M,m}(\alpha,\beta) = \mathcal{L}_{M,m}(\alpha',\beta')$.

(2) Let $\boldsymbol{f}$ be a maximizer of $\overline{\mathbf{L}^{\alpha,\beta}_{M,m}}$. Then $\boldsymbol{f}$ is also a maximizer of $\overline{\mathbf{L}^{\alpha',\beta'}_{M,m}}$.

Proof. (1) We first note π is feasible for $\mathcal{L}_{M,m}(\alpha',\beta')$ by construction, because $\pi^1_\#(A) \leq \alpha'(A)$ and $\pi^2_\#(A) \leq \beta'(A)$ for an arbitrary measurable set A, and $\pi(\Omega \times \Omega) \geq m$. Then we show π is the solution to $\mathcal{L}_{M,m}(\alpha',\beta')$ by contradiction. Assume there exists a transport plan π' that is better than π for $\mathcal{L}_{M,m}(\alpha',\beta')$. It is easy to see π' is a feasible transport plan for $\mathcal{L}_{M,m}(\alpha,\beta)$, because for an arbitrary measurable set A, $\pi'^1_\#(A) \leq \alpha'(A) \leq \alpha(A)$, and similarly, $\pi'^2_\#(A) \leq \beta(A)$. Thus π' is better than π for $\mathcal{L}_{M,m}(\alpha,\beta)$, which contradicts the optimality of π. Therefore, we conclude that π is the solution to $\mathcal{L}_{M,m}(\alpha',\beta')$, and the equality of the transport costs follows.

(2) First, we show that $\boldsymbol{f}$ is a maximizer of $\overline{\mathbf{L}^{\alpha',\beta}_{M,m}}$. We note $\boldsymbol{f}$ is in the feasible set of $\overline{\mathbf{L}^{\alpha',\beta}_{M,m}}$, thus we have

$$\begin{aligned}
\overline{\mathbf{L}^{\alpha',\beta}_{M,m}}(\boldsymbol{f}) &= \int_\Omega \boldsymbol{f} d\alpha' - \int_\Omega \boldsymbol{f} d\beta - \inf(\boldsymbol{f})(m - m_\beta) \\
&\geq \int_\Omega \boldsymbol{f} d\alpha - \int_\Omega \boldsymbol{f} d\beta - \inf(\boldsymbol{f})(m - m_\beta) \\
&= \overline{\mathbf{L}^{\alpha,\beta}_{M,m}}(\boldsymbol{f}) \\
&= \mathcal{L}_{M,m}(\alpha,\beta),
\end{aligned}$$

where the inequality holds because $\boldsymbol{f} \leq 0$ and $\alpha' \leq \alpha$. In addition, according to the first part of this proof, we have $\mathcal{L}_{M,m}(\alpha',\beta) = \mathcal{L}_{M,m}(\alpha,\beta)$, which implies $\overline{\mathbf{L}^{\alpha',\beta}_{M,m}}(\boldsymbol{f}) \geq \mathcal{L}_{M,m}(\alpha',\beta)$. Thus we conclude that $\boldsymbol{f}$ is a maximizer of $\overline{\mathbf{L}^{\alpha',\beta}_{M,m}}$.

Then, according to the symmetry stated in Proposition 14, $-\boldsymbol{f} + \inf(\boldsymbol{f})$ is a maximizer of $\overline{\mathbf{L}_{M,m}^{\beta,\alpha'}}$. By repeating the argument above, we claim $-\boldsymbol{f} + \inf(\boldsymbol{f})$ is a also maximizer of $\overline{\mathbf{L}_{M,m}^{\beta',\alpha'}}$. Finally, by using the symmetry again, we claim $\boldsymbol{f}$ is a maximizer of $\overline{\mathbf{L}_{M,m}^{\alpha',\beta'}}$. □

Proposition 15 has several important consequences. In particular, it implies the KR potentials are "flat" on the inactive mass. Formally, we have the following proposition.

Proposition 16 (Flatness of potentials)**.** *Let $\boldsymbol{f}$ be a maximizer of $\overline{\mathbf{L}_{M,m}^{\alpha,\beta}}$, and π be the solution to the primal form of $\mathcal{L}_{M,m}^{\alpha,\beta}$. Let $\alpha_{\mathcal{S}}$ and $\beta_{\mathcal{S}}$ be measures satisfying*

$$\begin{aligned} \pi_{\#}^{1} &\leq \alpha - \alpha_{\mathcal{S}} \leq \alpha \\ \textit{and} \quad \pi_{\#}^{2} &\leq \beta - \beta_{\mathcal{S}} \leq \beta \end{aligned}$$

Then $\boldsymbol{f} = 0$ $\alpha_{\mathcal{S}}$-almost surly, and $\boldsymbol{f} = -\inf(\boldsymbol{f})$ $\beta_{\mathcal{S}}$-almost surly.

Proof. According to Proposition 15, we claim that $\mathcal{L}_{M,m}^{\alpha-\alpha_{\mathcal{S}},\beta} = \mathcal{L}_{M,m}^{\alpha,\beta}$, and $\boldsymbol{f}$ is a maximizer of $\overline{\mathbf{L}_{M,m}^{\alpha-\alpha_{\mathcal{S}},\beta}}$. In other words, we have

$$\int_{\Omega} \boldsymbol{f} d\alpha - \int_{\Omega} \boldsymbol{f}\beta - \inf(\boldsymbol{f})(m - m_{\beta}) = \int_{\Omega} \boldsymbol{f} d(\alpha - \alpha_{\mathcal{S}}) - \int_{\Omega} \boldsymbol{f}\beta - \inf(\boldsymbol{f})(m - m_{\beta}).$$

By cleaning this equation, we obtain $\int_{\Omega} \boldsymbol{f} d\alpha_{\mathcal{S}} = 0$. Since $\boldsymbol{f} \leq 0$ and $\alpha_{\mathcal{S}}$ is a non-negative measure, we conclude that $\boldsymbol{f} = 0$ $\alpha_{\mathcal{S}}$-almost surely. The statement for $\beta_{\mathcal{S}}$ can be proved in a similar way. □

We can further characterize the inactive region where the mass is inactive. To begin with, we formally define the inactive region as follows

Definition 5 (Inactive region)**.** *A set A is inactive if it satisfies*

$$\pi(A \times \Omega) = 0 \quad \textit{and} \quad \pi(\Omega \times A) = 0 \tag{48}$$

where π is a solution to problem 16.

According to Corollary 1, as long as the optimal π transports mass from x to y, we have $\boldsymbol{f}(x) = \boldsymbol{f}(y) + d(x, y)$ almost surely. Thus, when a point x in α is not related to any point in β in this way, it must be inactive. Formally, we have the following proposition.

Proposition 17 (Inactive region)**.** *Let $\boldsymbol{f}$ denote a maximizer of $\overline{\mathbf{L}_{M,m}^{\alpha,\beta}}$. Denote set*

$$\mathcal{I}_{\alpha} = \Big\{ x | \boldsymbol{f}(x) < \boldsymbol{f}(y) + d(x, y), \ \forall y \in \textit{supp}(\beta) \Big\} \tag{49}$$

and

$$\mathcal{I}_{\beta} = \Big\{ y | \boldsymbol{f}(y) > \boldsymbol{f}(x) - d(x, y), \ \forall x \in \textit{supp}(\alpha) \Big\}. \tag{50}$$

Both $\mathcal{I}_{\alpha}$ and $\mathcal{I}_{\beta}$ are inactive.

Proof. We prove $\mathcal{I}_{\alpha}$ is inactivate by contradiction. If $\pi(\mathcal{I}_{\alpha} \times \Omega) > 0$, then $\pi(\mathcal{I}_{\alpha} \times \textit{supp}(\beta)) = \pi(\mathcal{I}_{\alpha} \times \Omega) > 0$, because $\pi(\mathcal{I}_{\alpha} \times (\Omega \backslash \textit{supp}(\beta))) \leq \pi(\Omega \times (\Omega \backslash \textit{supp}(\beta))) \leq \beta(\Omega \backslash \textit{supp}(\beta)) = 0$. As a result, for arbitrarily small $\epsilon > 0$, there exists a compact set $K \subseteq \mathcal{I}_{\alpha} \times \textit{supp}(\beta)$ such that $\pi(K) > \pi(\mathcal{I}_{\alpha} \times \textit{supp}(\beta)) - \epsilon > 0$. Since $d(x, y) - \boldsymbol{f}(x) + \boldsymbol{f}(y)$ is continuous and K is compact, there exists $(\widetilde{x}, \widetilde{y}) \in K \subseteq \mathcal{I}_{\alpha} \times \textit{supp}(\beta)$ such that

$$\inf_{(x,y) \in K} (d(x, y) - \boldsymbol{f}(x) + \boldsymbol{f}(y)) = d(\widetilde{x}, \widetilde{y}) - \boldsymbol{f}(\widetilde{x}) + \boldsymbol{f}(\widetilde{y}) > 0.$$

Note the last inequality holds by the definition of $\mathcal{I}_\alpha$. In addition, we have $d(x,y) - \boldsymbol{f}(x) + \boldsymbol{f}(y) \geq 0$ for all $(x,y) \in \Omega \times \Omega$ because $\boldsymbol{f} \in Lip(\Omega)$. By combining these inequalities, we have

$$\begin{aligned}
&\int_{\Omega\times\Omega} (d(x,y) - \boldsymbol{f}(x) + \boldsymbol{f}(y)) d\pi(x,y) \\
&\geq \int_{K} (d(x,y) - \boldsymbol{f}(x) + \boldsymbol{f}(y)) d\pi(x,y) \\
&\geq \inf_{(x,y)\in K} (d(x,y) - \boldsymbol{f}(x) + \boldsymbol{f}(y))(\pi(\mathcal{I}_\alpha \times \Omega) - \epsilon) \\
&> 0.
\end{aligned}$$

However, this contradicts Corollary 1, which claims that

$$\int_{\Omega\times\Omega} (d(x,y) - \boldsymbol{f}(x) + \boldsymbol{f}(y)) d\pi(x,y) = 0.$$

We thus conclude $\pi(\mathcal{I}_\alpha \times \Omega) = 0$. In addition, it is easy to check $\mathcal{I}_\alpha \subseteq \Omega \backslash supp(\beta)$, since for all $x \in supp(\beta)$, there exists $y = x$ such that $\boldsymbol{f}(x) = \boldsymbol{f}(y) + d(x,y)$, which suggests $x \notin \mathcal{I}_\alpha$. Thus $\pi(\Omega \times \mathcal{I}_\alpha) \leq \beta(\mathcal{I}_\alpha) \leq \beta(\Omega \backslash supp(\beta)) = 0$. Finally, we can finally conclude that $\mathcal{I}_\alpha$ is inactive.

We omit the proof of $\mathcal{I}_\beta$ since it can be proved in a similar way. □

Based on Proposition 17, we can construct an inactive region as follows.

Corollary 3 (Boundary of active region)**.** *Let $\boldsymbol{f}$ denote a maximizer of $\overline{\mathbf{L}_{M,m}^{\alpha,\beta}}$. Denote*

$$\mathcal{F}_\alpha = \left\{ x | dist(x, supp(\beta)) > -\inf(\boldsymbol{f}) \right\}$$

and

$$\mathcal{F}_\beta = \left\{ y | dist(y, supp(\alpha)) > -\inf(\boldsymbol{f}) \right\}.$$

Both $\mathcal{F}_\alpha$ and $\mathcal{F}_\beta$ are inactive.

Proof. For arbitrary $x \in \mathcal{F}_\alpha$ and arbitrary $y \in supp(\beta)$, we have $\boldsymbol{f}(y) + d(x,y) > \boldsymbol{f}(y) - \inf(\boldsymbol{f}) \geq 0 \geq \boldsymbol{f}(x)$, thus $x \in \mathcal{I}_\alpha$. Therefore, $\mathcal{F}_\alpha \subseteq \mathcal{I}_\alpha$. According to Proposition 17, $\mathcal{I}_\alpha$ is inactive, thus so is $\mathcal{F}_\alpha$. We omit the proof of $\mathcal{F}_\beta$ since it can be proved in a similar way. □

Intuitively, $\boldsymbol{f}$ is flat in the inactive area described in Corollary 3. This is formally stated as follows.

Corollary 4 (Flat region of potentials)**.** *Let $\boldsymbol{f}$ be a maximizer of $\overline{\mathbf{L}_{M,m}^{\alpha,\beta}}$. Define $\mathcal{F}_\alpha$ and $\mathcal{F}_\beta$ as in Corollary 3, and the corresponding measures $\alpha_{\mathcal{F}}$ and $\beta_{\mathcal{F}}$ as*

$$\begin{aligned}
\alpha_{\mathcal{F}}(A) &= \alpha(A \cap \mathcal{F}_\alpha) \\
\textit{and} \quad \beta_{\mathcal{F}}(A) &= \beta(A \cap \mathcal{F}_\beta)
\end{aligned}$$

for an arbitrary measurable set A. Then $\boldsymbol{f} = 0$ $\alpha_{\mathcal{F}}$-almost surly, and $\boldsymbol{f} = -\inf(\boldsymbol{f})$ $\beta_{\mathcal{F}}$-almost surly.

Proof. We only prove the statement for $\alpha_{\mathcal{F}}$ here and the statement for $\beta_{\mathcal{F}}$ can be proved similarly. According to Proposition 16, we only need to prove $\pi^1_\# \leq \alpha - \alpha_{\mathcal{F}} \leq \alpha$. Since the second inequality holds as $\alpha_{\mathcal{F}} \geq 0$, we only need to prove the first inequality. In other words, we need to prove

$$\pi^1_\#(A) + \alpha(A \cap \mathcal{F}_\alpha) \leq \alpha(A)$$

for an arbitrary measurable set A. For the first term in the left hand side, we have $\pi^1_\#(A) = \pi^1_\#(A \cap supp(\pi^1_\#)) \leq \alpha(A \cap supp(\pi^1_\#))$. In addition, since $\pi^1_\#(\mathcal{F}_\alpha) = 0$ according to Corollary 3, and $\mathcal{F}_\alpha$ is open, we have $\mathcal{F}_\alpha \cap supp(\pi^1_\#) = \varnothing$. Therefore, $(A \cap supp(\pi^1_\#)) \cap (A \cap \mathcal{F}_\alpha) = \varnothing$, thus $\alpha(A \cap supp(\pi^1_\#)) + \alpha(A \cap \mathcal{F}_\alpha) = \alpha((A \cap supp(\pi^1_\#)) \cup (A \cap \mathcal{F}_\alpha) \leq \alpha(A)$, which finishes the proof. □

Another consequence of Proposition 15 is that it implies $\mathcal{L}_{M,m}$ actually solves a $\mathcal{W}_1$ problem on the active measure. This is formally stated in the following proposition.

Proposition 18 (Reduce to $\mathcal{W}_1$)**.** *Let $\boldsymbol{f}$ be a maximizer of $\mathbf{L}_{M,m}$, and π be the solution to the corresponding primal form of $\mathcal{L}_{M,m}$ 16.*

(1) π is the solution to the primal form of $\mathcal{W}_1(\pi^1_\#, \pi^2_\#)$.

(2) $\boldsymbol{f}$ is a maximizer of $\mathbf{W}_1(\pi^1_\#, \pi^2_\#)$.

Proof. First, we note that π only transport the necessary amount of mass, *i.e.*, $m = \pi^1_\#(\Omega) = \pi^2_\#(\Omega) = \pi(\Omega \times \Omega)$, because if $m < \pi(\Omega \times \Omega)$, we obtain a better solution $\pi' = \frac{m}{\pi(\Omega\times\Omega)}\pi$, which contradicts the optimality of π. Then let $\alpha' = \pi^1_\#$ and $\beta' = \pi^2_\#$, Proposition 15 suggests that π is respectively the solution for the primal forms $\mathcal{L}_{M,m}(\pi^1_\#, \pi^2_\#) = \mathcal{W}_1(\pi^1_\#, \pi^2_\#)$. In addition, since $m = m_{\beta'}$, Proposition 15 suggests $\boldsymbol{f}$ is a maximizer of $\mathbf{W}^{\alpha',\beta'} = \overline{\mathbf{L}^{\alpha',\beta'}_{M,m}}$, which finishes the proof. □

In summary, Proposition 16 and 18 provide an intuitive characterization for $\mathcal{L}_{M,m}$. Specifically, solving $\mathcal{L}_{M,m}$ can be regarded as solving two sub-problems. The first sub-problem is to identify the active measures and the inactive measures. The second sub-problem is to perform the complete optimal transport between the active measures, and to discard the inactive measures. In this sense, Corollary 2 becomes straightforward, because when both α and β are fully active, $\mathcal{L}_{M,m}(\alpha, \beta)$ simply becomes the second sub-problem, *i.e.*, $\mathcal{W}_1(\alpha, \beta)$.

D.4.2 Distance-type PW discrepancy

Now we discuss the properties of $\mathcal{L}_{D,h}$ as analogues of those of $\mathcal{L}_{M,m}$. We only present the results and omit the proof, since they can be proved in a very similar manner to their counterparts in the previous subsection.

Proposition 19. *Let $\boldsymbol{f}$ denote a maximizer to $\mathbf{L}^{\alpha,\beta}_{D,h}$. If $m_\alpha > m_\beta$, then $\sup(\boldsymbol{f}) = 0$; If $m_\beta > m_\alpha$, then $\inf(\boldsymbol{f}) = -h$.*

Proposition 20 (Symmetry)**.** *Let $\boldsymbol{f}$ denote a maximizer to $\mathbf{L}^{\alpha,\beta}_{D,h}$. $\tilde{\boldsymbol{f}} = -\boldsymbol{f} - h$ is a maximizer of $\mathbf{L}^{\beta,\alpha}_{D,h}$. In addition, we have $\mathcal{L}_{D,h}(\alpha, \beta) = \mathcal{L}_{D,h}(\beta, \alpha)$.*

Proposition 21 (Inactive mass)**.** *Let π be the solution to the primal form of $\mathcal{L}_{D,h}(\alpha, \beta)$. Define α' and β' be measures satisfying*

$$\pi^1_\#(A) \le \alpha'(A) \le \alpha(A) \quad \textit{and} \quad \pi^2_\#(A) \le \beta'(A) \le \beta(A) \tag{51}$$

for an arbitrary measurable set A.

(1) π is also the solution to the primal form of $\mathcal{L}_{D,h}(\alpha', \beta')$, thus $\mathcal{L}_{D,h}(\alpha, \beta) = \mathcal{L}_{D,h}(\alpha', \beta')$.

(2) Let $\boldsymbol{f}$ be a maximizer of $\mathbf{L}^{\alpha,\beta}_{D,h}$. Then $\boldsymbol{f}$ is also a maximizer of $\mathbf{L}^{\alpha',\beta'}_{D,h}$.

Proposition 22 (Flatness of potentials)**.** *Let $\boldsymbol{f}$ be a maximizer of $\mathbf{L}^{\alpha,\beta}_{D,h}$, and π be the solution to the primal form of $\mathcal{L}^{\alpha,\beta}_{D,h}$. Let $\alpha_{\mathcal{S}}$ and $\beta_{\mathcal{S}}$ be measures satisfying*

$$\begin{aligned} \pi^1_\# &\le \alpha - \alpha_{\mathcal{S}} \le \alpha \\ \textit{and} \quad \pi^2_\# &\le \beta - \beta_{\mathcal{S}} \le \beta. \end{aligned}$$

Then $\boldsymbol{f} = 0$ $\alpha_{\mathcal{S}}$-almost surly, and $\boldsymbol{f} = -h$ $\beta_{\mathcal{S}}$-almost surly.

Proposition 23 (Inactive region)**.** *Let $\boldsymbol{f}$ denote a maximizer to $\mathbf{L}^{\alpha,\beta}_{D,h}$. Denote set*

$$\mathcal{I}_\alpha = \Big\{ x | \boldsymbol{f}(x) < \boldsymbol{f}(y) + d(x, y),\ \forall y \in \textit{supp}(\beta) \Big\} \tag{52}$$

and

$$\mathcal{I}_\beta = \Big\{ y | \boldsymbol{f}(y) > \boldsymbol{f}(x) - d(x, y),\ \forall x \in \textit{supp}(\alpha) \Big\}. \tag{53}$$

Both $\mathcal{I}_\alpha$ and $\mathcal{I}_\beta$ are inactive.

Corollary 5 (Boundary of active region)**.** *For $\mathcal{L}_{D,h}$, denote*

$$\mathcal{F}_\alpha = \Big\{x | dist(x, supp(\beta)) > h\Big\} \tag{54}$$

and

$$\mathcal{F}_\beta = \Big\{y | dist(y, supp(\alpha)) > h\Big\}. \tag{55}$$

Both $\mathcal{F}_\alpha$ and $\mathcal{F}_\beta$ are inactive.

Corollary 6 (Flat region of potentials)**.** *Let $\boldsymbol{f}$ be a maximizer of $\mathbf{L}_{D,h}^{\alpha,\beta}$. Define $\mathcal{F}_\alpha$ and $\mathcal{F}_\beta$ as in Corollary 5, and the corresponding measures $\alpha_{\mathcal{F}}$ and $\beta_{\mathcal{F}}$ as*

$$\begin{aligned} \alpha_{\mathcal{F}}(A) &= \alpha(A \cap \mathcal{F}_\alpha) \\ \textit{and} \quad \beta_{\mathcal{F}}(A) &= \beta(A \cap \mathcal{F}_\beta) \end{aligned}$$

for an arbitrary measurable set A. Then $\boldsymbol{f} = 0$ $\alpha_{\mathcal{F}}$-almost surly, and $\boldsymbol{f} = -h$ $\beta_{\mathcal{F}}$-almost surly.

Proposition 24 (Reduce to $\mathcal{W}_1$)**.** *Let $\boldsymbol{f}$ be a maximizer of $\mathbf{L}_{D,h}$, and π be the solution to the corresponding primal form of $\mathcal{L}_{D,h}$ 21.*

(1) π is the solution to the primal form of $\mathcal{W}_1(\pi^1_\#, \pi^2_\#)$.

(2) $\boldsymbol{f}$ is a maximizer of $\mathbf{W}_1(\pi^1_\#, \pi^2_\#)$.

D.5 Differentiability

Now we consider the differentiability of the PW discrepancy. We show that, when expressed in the KR forms, if a mild assumption is satisfied, then $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$ are differentiable and the gradients can be explicitly computed, which makes it feasible to optimize them via gradient based methods.

We consider a parametrized transformation $\mathcal{T}_\theta : \mathbb{R}^d \times \Omega \to \Omega$, and denote β_θ the corresponding push-forward measure of β. We show in Proposition 25 and 26 that, if $\mathcal{T}_\theta$ is Lipschitz w.r.t. θ, then the objective functions $\mathcal{L}_{D,h}(\alpha, \beta_\theta)$ and $\mathcal{L}_{M,m}(\alpha, \beta_\theta)$ is differentiable w.r.t. θ, and the gradient can be computed explicitly.

Proposition 25 (Differentiability of $\mathcal{L}_{D,h}$)**.** *If $\mathcal{T}_\theta : \Omega \to \Omega$ is Lipschitz w.r.t. to θ, then $\mathcal{L}_{D,h}(\alpha, \beta_\theta)$ is continuous w.r.t. θ, and is differentiable almost everywhere. Furthermore, we have*

$$\nabla_\theta \mathcal{L}_{D,h}(\alpha, \beta_\theta) = -\int_\Omega \nabla_\theta \boldsymbol{f}(\mathcal{T}_\theta(x)) d\beta, \tag{56}$$

where $\boldsymbol{f}$ is a maximizer of $\mathbf{L}_{D,h}$.

Proof. We first prove $\mathcal{L}_{D,h}(\alpha, \beta_\theta)$ is continuous w.r.t. θ. By definition, we have $\mathcal{L}_{D,h}(\alpha, \beta) = KR_h(\alpha, \beta) - \frac{h}{2}(m_\alpha + m_\beta)$. Thus by triangle inequality of the KR metric, for arbitrary $\theta, \theta' \in \mathbb{R}^d$, we have

$$\begin{aligned} &|\mathcal{L}_{D,h}(\alpha, \beta_\theta) - \mathcal{L}_{D,h}(\alpha, \beta_{\theta'})| \\ =&|\Big(KR_h(\alpha, \beta_\theta) - \frac{h}{2}(m_\alpha + m_{\beta_\theta})\Big) - \Big(KR_h(\alpha, \beta_{\theta'}) - \frac{h}{2}(m_\alpha + m_{\beta_{\theta'}})\Big)| \\ =&|KR_h(\alpha, \beta_\theta) - KR_h(\alpha, \beta_{\theta'})| \\ \leq& KR_h(\beta_{\theta'}, \beta_\theta) \end{aligned} \tag{57}$$

Note that the second equality holds because $\mathcal{T}$ maintains the total mass, *i.e.*, $m_{\beta_{\theta'}} = m_{\beta_\theta} = m_\beta$. In addition, we have

$$KR_h(\beta_{\theta'}, \beta_\theta) \leq \mathcal{W}_1(\beta_{\theta'}, \beta_\theta) \leq \int_\Omega d(\mathcal{T}_\theta(y) - \mathcal{T}_{\theta'}(y)) d\beta(y).$$

By assumption, $\mathcal{T}_\theta$ is Lipschitz w.r.t. θ, thus there exists a constant $\Delta > 0$, such that $d(\mathcal{T}_\theta(y), \mathcal{T}_{\theta'}(y)) \leq \Delta ||\theta - \theta'||$. Thus $KR_h(\beta_{\theta'}, \beta_\theta) \leq m_\beta \Delta ||\theta - \theta'||$. Finally, we have

$$\mathcal{L}_{D,h}(\alpha, \beta_\theta) - \mathcal{L}_{D,h}(\alpha, \beta_{\theta'}) \leq KR_h(\beta_{\theta'}, \beta_\theta) \leq m_\beta \Delta ||\theta - \theta'||,$$

which proves $\mathcal{L}_{D,h}(\alpha, \beta_\theta)$ is continuous w.r.t. θ, and Radamacher's theorem states that it is differentiable almost everywhere.

The rest of the proof follows that in Arjovsky et al. (2017). The sketch of the proof is as follows. We first derive $\nabla_\theta \mathcal{L}_{D,h}(\alpha, \beta_\theta)$ as $-\nabla_\theta \int_\Omega \boldsymbol{f}(\mathcal{T}_\theta(x))d\beta$ using the envelope theorem (Milgrom & Segal, 2002). Then we prove

$$\nabla_\theta \int_\Omega \boldsymbol{f}(\mathcal{T}_\theta(x))d\beta = \int_\Omega \nabla_\theta \boldsymbol{f}(\mathcal{T}_\theta(x))d\beta,$$

when the right hand side of the equation is well defined, which completes our proof. □

Proposition 26 (Differentiability of $\mathcal{L}_{M,m}$)**.** *If $\mathcal{T}_\theta : \Omega \to \Omega$ is Lipschitz w.r.t. to θ, then $\mathcal{L}_{M,m}(\alpha, \beta_\theta)$ is continuous w.r.t. θ, and is differentiable almost everywhere. Furthermore, we have*

$$\nabla_\theta \mathcal{L}_{M,m}(\alpha, \beta_\theta) = -\int_\Omega \nabla_\theta \boldsymbol{f}(\mathcal{T}_\theta(x))d\beta, \tag{58}$$

where $(\boldsymbol{f}, h)$ is a maximizer of $\mathbf{L}_{M,m}$.

Proof. To begin with, for arbitrary θ, θ', we consider the following two-step procedure that transports m unit mass from α to $\beta_{\theta'}$. First, we transport m mass from α to β_θ according to $\overline{\pi}$, which is a solution to $\mathcal{L}_{M,m}(\alpha, \beta_\theta)$. Second, we transport m mass received in β_θ to $\beta_{\theta'}$ according to a plan π' defined as

$$\pi'(A \times B) = \begin{cases} \widetilde{\pi}(A \times B)\frac{\overline{\pi}^2_\#(A)}{\beta_\theta(A)} & if\ \beta_\theta(A) > 0 \\ 0 & else, \end{cases}$$

where $\widetilde{\pi}(\mathcal{T}_\theta(x), \mathcal{T}_{\theta'}(x)) = \beta(x)$. Thus we have

$$\int_{\Omega\times\Omega} d(x,z)d\pi'(x,z) \leq \int_{\Omega\times\Omega} d(x,z)d\widetilde{\pi}(x,z) = \int_\Omega d(\mathcal{T}_\theta(x), \mathcal{T}_{\theta'}(x))d\beta(x).$$

Since $\mathcal{T}_\theta$ is Lipschitz w.r.t. θ, there exists a constant $\Delta > 0$, such that $d(\mathcal{T}_\theta(x), \mathcal{T}_{\theta'}(x)) \leq \Delta||\theta - \theta'||$. Therefore, the cost of the second step can be bounded as

$$\int_{\Omega\times\Omega} d(x,z)d\pi'(x,z) \leq m_\beta \Delta||\theta - \theta'||.$$

Denote $cost_1$ the overall cost of this two-step procedure. We have

$$cost_1 \leq \mathcal{L}_{M,m}(\alpha, \beta_\theta) + m_\beta \Delta||\theta - \theta'||.$$

By applying the gluing lemma (Villani (2009) Sec.1) to $\overline{\pi}$ and π', we can construct a transport plan $\widetilde{\pi}$ that transports m unit mass from $\overline{\pi}^1_\#$ to $\pi'^2_\#$. Denote $cost_2 = \int_{\Omega\times\Omega} d(x,y)d\widetilde{\pi}(x,y)$ the cost of $\widetilde{\pi}$. One one hand, we have

$$\mathcal{L}_{M,m}(\alpha, \beta_{\theta'}) \leq cost_2$$

according to the definition of $\mathcal{L}_{M,m}$. On the other hand, we have

$$cost_2 \leq cost_1.$$

Because for arbitrary $x, z \in \Omega$, the two-step procedure and $\widetilde{\pi}$ transport the same amount of mass between them. However, the cost of transporting a unit mass from x to z is $d(x,y) + d(y,z)$ for a $y \in \Omega$ for the two-step procedure, but is $d(x,z)$ for $\widetilde{\pi}$, which is cheaper according to triangle inequality. By combining these inequalities together, we have

$$\mathcal{L}_{M,m}(\alpha, \beta_{\theta'}) \leq cost_2 \leq cost_1 \leq \mathcal{L}_{M,m}(\alpha, \beta_\theta) + m_\beta \Delta||\theta - \theta'||,$$

i.e.,

$$\mathcal{L}_{M,m}(\alpha, \beta_{\theta'}) \leq \mathcal{L}_{M,m}(\alpha, \beta_\theta) + m_\beta \Delta||\theta - \theta'||.$$

By switching θ and θ' and repeating the argument, we have

$$\mathcal{L}_{M,m}(\alpha, \beta_\theta) \leq \mathcal{L}_{M,m}(\alpha, \beta_{\theta'}) + m_\beta \Delta||\theta - \theta'||,$$

thus we have

$$|\mathcal{L}_{M,m}(\alpha, \beta_\theta) - \mathcal{L}_{M,m}(\alpha, \beta_{\theta'})| \leq m_\beta \Delta||\theta - \theta'||,$$

which suggests $\mathcal{L}_{M,m}(\alpha, \beta_\theta)$ is continuous w.r.t. θ, and Radamacher's theorem states that it is differentiable almost everywhere. The rest of the proof is similar to that of Proposition 25. □

Remark The main idea used in the proofs in this section is based on that in Arjovsky et al. (2017). However, a key difference is that unlike $\mathcal{W}_1$, neither $\mathcal{L}_{M,m}$ nor $\mathcal{L}_{D,h}$ is a metric, *i.e.*, they do not necessarily satisfy triangle inequality, thus the differences cannot be bounded directly. This gap is bridged by using the KR metric in Proposition 25, and by constructing a transportation plan in Proposition 26.

E APPLICATION TO POINT REGISTRATION

In this section, we focus on the discrete distributions and prove the main results in the main text.

We first present the proof of the main Theorems in the main text.

Proof of Theorem 1. The KR formulation of $\mathcal{L}_{D,h}$ is given in Proposition 1, and the existence of the optimizer is proved in Proposition 2. The gradient of $\mathcal{L}_{D,h}(\alpha, \beta_\theta)$ is derived in Proposition 25. Theorem 1 is proved by applying these results to discrete distributions. □

Proof of Theorem 2. The KR formulation of $\mathcal{L}_{M,m}$ is given in Theorem 3, and the existence of the optimizer is proved in Proposition 5. The gradient of $\mathcal{L}_{M,m}(\alpha, \beta_\theta)$ is derived in Proposition 26. Theorem 2 is proved by applying these results to discrete distributions. □

We finally verify that the parametrized transformation $\mathcal{T}_\theta$ used in our paper satisfies the regularity assumption, *i.e.*, it is Lipschitz w.r.t. the parameter θ.

Lemma 5. *If $\mathcal{T} : \mathbb{R}^d \times \Omega \to \Omega$ is Lipschitz w.r.t. to its first variable, then for arbitrary $y \in Y$, and θ, $\theta' \in \mathbb{R}^d$, there exists a $\Delta > 0$, such that*

$$||\mathcal{T}_\theta(y) - \mathcal{T}_{\theta'}(y)|| \leq \Delta ||\theta - \theta'||.$$

Proof. The proof is straightforward. By assumption, for arbitrary $y_i \in Y$, we have $||\mathcal{T}_\theta(y_i) - \mathcal{T}_{\theta'}(y_i)|| \leq \Delta_i ||\theta - \theta'||$. Since Y is finite, we can simply let $\Delta = \max_i(\{\Delta_i\}_{y_i \in Y})$. □

Proposition 27. *Given a point set $Y = \{y_i\}_{i=1}^r \subseteq \Omega$. Let $\theta = (A, t, V) \in \mathbb{R}^{12+3r}$, where $A \in \mathbb{R}^{3\times 3}$ is a affinity matrix, $t \in \mathbb{R}^{1\times 3}$ is a translation vector, $V \in \mathbb{R}^{r\times 3}$ is the offset vectors for all points in Y, and V_i represent the i-th row of V. For each point $y_i \in Y$, define $\mathcal{T}_\theta(y_i) = y_i A + t + V_i$, where V_i represents the i-th row of V. $\mathcal{T} : \mathbb{R}^d \times \Omega \to \Omega$ is Lipschitz w.r.t. to its first variable.*

Proof. Note for arbitrary A, there exists $\delta > 0$, such that $||A||_2 \leq \delta ||A||_F$, where $||\cdot||_F$ is the Frobenius norm. For arbitrary $\theta = (A, t, V)$, $\widetilde{\theta} = (\widetilde{A}, \widetilde{t}, \widetilde{V})$ and $y_i \in Y$, we have

$$\begin{aligned}
||\mathcal{T}_\theta(y_i) - \mathcal{T}_{\widetilde{\theta}}(y_i)||_2 &= ||(A - \widetilde{A})y_i + (t - \widetilde{t}) + (V_i - \widetilde{V}_i)||_2 \\
&\leq ||A - \widetilde{A}||_2 ||y_i||_2 + ||t - \widetilde{t}||_2 + ||V_i - \widetilde{V}_i||_2 \\
&\leq \delta ||A - \widetilde{A}||_F ||y_i||_2 + ||\theta - \widetilde{\theta}||_2 + ||\theta - \widetilde{\theta}||_2 \\
&\leq \delta ||\theta - \widetilde{\theta}||_2 ||y_i||_2 + ||\theta - \widetilde{\theta}||_2 + ||\theta - \widetilde{\theta}||_2 \\
&= (\delta ||y_i|| + 2)||\theta - \widetilde{\theta}||_2,
\end{aligned}$$

which proves that $\mathcal{T}$ is Lipschitz w.r.t. to its first variable. □

F MORE DETAILS OF THE MAIN TEXT

F.1 MORE DETAILS OF SEC. 2

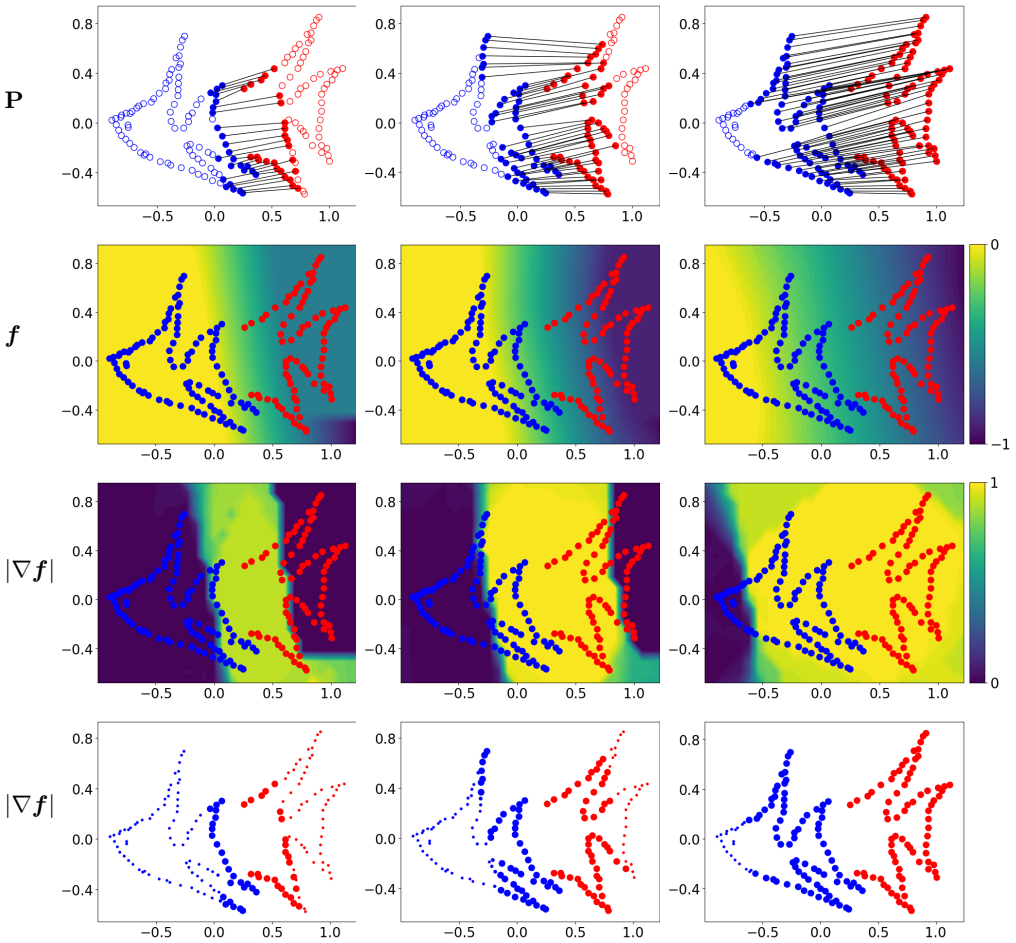
In contrast to the probabilistic methods, the correspondence-based methods explicitly estimate the correspondence between point sets, and then compute the transformation accordingly. The iterative closest point (ICP) algorithm (Besl & McKay, 1992; Zhang, 1994) is one of the well-known methods in this class. In each iteration, ICP estimates the correspondence based on the nearest distance criterion, and updates the transformation accordingly. Despite its simplicity, ICP easily fails when the initial point sets are not roughly aligned. In order to improve the robustness of ICP, several

methods (Fitzgibbon, 2003; Granger & Pennec, 2002; Chui & Rangarajan, 2000; Stewart et al., 2003) have been proposed. In particular, Chui & Rangarajan (2000) proposed to estimate the correspondence between all pairs of points, which improves the robustness at the expense of high computational cost.

Our method is also related to Wasserstein generative network (WGAN) (Arjovsky et al., 2017), which is a popular method for large scale DM problems. A recent survey of the applications of WGAN can be found in Gui et al. (2020). WGAN efficiently optimizes the Wasserstein-1 metric by approximating Wasserstein-1 potential using a neural network. This technique is also used in our method to efficiently optimize the PW discrepancy. In this sense, our method directly generalizes WGAN to PDM problems.

F.2 More details of Sec. 4.2

We present a simple example comparing the primal and the KR forms in Fig. 9. We show the correspondence $\mathbf{P}$ obtained by the primal form in the 1-st row, where the black lines link the corresponding points. To estimate the KR forms, we parametrize the potential function $\boldsymbol{f}_{w,h}$ by a neural network shown in Fig. 11, and learn $\boldsymbol{f}_{w,h}$ by maximizing (9) or (10). The 2-nd row visualizes the learned potential function $\boldsymbol{f}_{w,h}$. The corresponding gradient norms $|\nabla \boldsymbol{f}_{w,h}|$ are shown in the 3-rd row. $|\nabla \boldsymbol{f}_{w,h}|$ are further visualized in the 4-th row, where the size of each point is approximately proportional to the gradient norm.



(a) $\mathcal{L}_{M,25}$ or $\mathcal{L}_{D,0.64}$ (b) $\mathcal{L}_{M,50}$ or $\mathcal{L}_{D,1.09}$ (c) $\mathcal{L}_{M,78}$ or $\mathcal{L}_{D,5}$

Figure 9: Comparison between the primal and the KR forms of the PW discrepancies. We consider the reference set α (blue) and the source set β (red). See text for details.

We further evaluate the precision of the estimated KR forms. For each setting, we compute $\mathcal{L}_{M,m}$ or $\mathcal{L}_{D,h}$ in the KR forms by inserting the learned potential $\boldsymbol{f}_{w,h}$ back into (9) or (10) (without the gradient penalty term). To obtain the true values of $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$, we compute their primal forms (1) and (2) using the POT tool box (Flamary et al., 2021). The results are summarized in Tab. 2. As can be seen, our estimated PW discrepancies are close to the true values with the averaged relative error less than 0.2%, which we think is sufficiently precise for our application.

As for the efficiency, we note that the KR formulation is not advantageous in this data scale, *i.e.*, tens of points, where the computation of the exact primal form only takes less than 1 sec on a CPU, while the computation of the KR form takes a few minutes on a mid-end GPU (due to the requirement of learning the potential network). However, for large scale dataset, *e.g.*, $\sim 10^6$, which the primal form can not handle because the transport map (of size $10^6 \times 10^6$) does not even fit into the memory, the KR form can still apply (as validate in our experiments) using a 12G GPU.

Finally, we stress that since the implement of KR form depends on the underlying neural network, both the efficiency and precision of the KR forms naturally vary according to the structure of the neural network. This is fundamentally different from the solvers of the primal form, such as the Sinkhorn algorithm or the linear program, which are fixed pipelines. Therefore, it is generally not possible to compare the efficiency and precision between the KR solvers and the primal solvers in a more rigorous way.

Table 2: Precision of the estimated PW discrepancies for the "fish" shape in Fig. 9

	$\mathcal{L}_{M,25}$	$\mathcal{L}_{M,50}$	$\mathcal{L}_{M,78}$	$\mathcal{L}_{D,0.648}$	$\mathcal{L}_{D,1.09}$	$\mathcal{L}_{D,5}$
True value	0.1354	0.4191	0.8955	-0.0722	-0.2795	-4.1044
KR (Ours)	0.1352	0.4202	0.8994	-0.0724	-0.2791	-4.1004

F.3 More details of Sec. 4.3

To estimate the gradient of the coherence energy fast, we first decompose $\mathbf{G}$ as $\mathbf{G} \approx \mathbf{Q}\Lambda\mathbf{Q}^T$ via the Nyström method (Williams & Seeger, 2000), where $k \ll r$, $\mathbf{Q} \in \mathbb{R}^{r\times k}$, and $\Lambda \in \mathbb{R}^{k\times k}$ is a diagonal matrix. Then we apply the Woodbury identity to $(\sigma\mathcal{I} + \mathbf{Q}\Lambda\mathbf{Q}^T)^{-1}$ and obtain

$$(\sigma\mathcal{I} + \mathbf{Q}\Lambda\mathbf{Q}^T)^{-1} = \sigma^{-1}\mathcal{I} - \sigma^{-2}\mathbf{Q}(\Lambda^{-1} + \sigma^{-1}\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T.$$

As a result, the gradient of the coherence energy can be approximated as

$$\frac{\partial\mathcal{C}(\mathcal{T}_\theta)}{\partial V} = 2\lambda(\sigma\mathcal{I} + \mathbf{G})^{-1}V \approx (2\lambda)(\sigma^{-1}V - \sigma^{-2}\mathbf{Q}(\Lambda^{-1} + \sigma^{-1}\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^TV).$$

F.4 Connections between PWAN and WGAN

In theory, PWAN includes WGAN as a special case. Because in the DM setting where distributions are required to be completely matched, *i.e.*, when $m_\alpha = m_\beta = m$ for $\mathcal{L}_{M,m}$, or $m_\alpha = m_\beta$ and $h > diam(\Omega)$ for $\mathcal{L}_{D,h}$, both $\mathcal{L}_{M,m}$ and $\mathcal{L}_{D,h}$ become $\mathcal{W}_1$ (Corollary 2), thus the bound constraint of the potential is not needed, *i.e.*, PWAN becomes WGAN. In practice, these two methods are only different in the structures of their potential networks. This is summarized in Tab. 3.

Table 3: Comparison between the potential networks of WGAN and PWAN.

	input threshold	Lipschitz	negative & bounded	lower bound h
WGAN	–	✓	–	–
d-PWAN	distance	✓	✓	fixed
m-PWAN	mass	✓	✓	learnable

To provide a more in-depth characterization of the relations between PWAN and WGAN, we consider the toy 1-dimensional example in Fig. 4 again. We compare the potentials of WGAN and PWAN in Fig. 10. As can be seen, both methods learn Lipschitz potentials. As discussed in the main text, the PWAN potential permits the existence of some flat regions where gradient norm is 0. In contrast, the WGAN potential has gradient norm 1 almost surely (Corollary 1 in Gulrajani et al. (2017)), *i.e.*, no

flat region exists. This difference explains the different behaviors of these two methods. Specifically, PWAN is a PDM method, because it will discard a fraction of points in β_θ within in flat region, while only moving the points outside the flat region toward α (as discussed in the main text). However, WGAN is a DM method, because it will move ALL points in β_θ toward α.

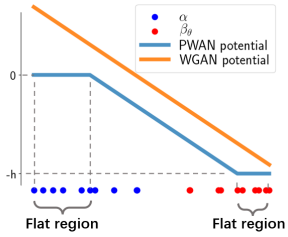


Figure 10: Comparison between the potentials learned by PWAN and WGAN.

We note that PWAN has the following adversarial explanation. Let us call the points in α real points, and those in β_θ fake points. In the registration process, $\boldsymbol{f}_{w,h}$ is trying to discriminate real and fake points by assigning each of them a score in range $[-h, 0]$, where real points have higher scores than fake points. $\boldsymbol{f}_{w,h}$ is so certain of a fraction of points, that it assigns the highest or lowest possible score (0 or $-h$) to them and does not change these scores easily. Meanwhile, $\mathcal{T}_\theta$ is trying to cheat $\boldsymbol{f}_{w,h}$ by moving the fraction of fake points of which $\boldsymbol{f}_{w,h}$ is not certain to obtain higher scores. In addition, $\mathcal{T}_\theta$ is also trying to move all fake points to keep the coherence energy low. The process ends when $\mathcal{T}_\theta$ cannot make further improvement on the fake points.

F.5 Experimental Details

Data synthesis The data synthesis procedure mostly follows Hirose (2021b). Specifically, given the original point set $X \in \mathbb{R}^{r\times 3}$ the deformed point set $Y \in \mathbb{R}^{r\times 3}$ is generated by:

$$Y = X + V + \epsilon, \tag{59}$$

where ϵ is the Gaussian noise following $\mathcal{N}(0, 0.02)$, and $V \in \mathbb{R}^{r\times 3}$ is the random offset vectors. To ensure V varies smoothly in the space, we sample column vector V_j from a distribution p:

$$p(v) = \mathcal{N}(v; 0, \lambda^{-1}\mathbf{G}), \tag{60}$$

where $\mathbf{G} \in \mathbb{R}^{r\times r}$ is a Gaussian kernel defined as $\mathbf{G}_\rho(i,j) = e^{-||y_i - y_j||^2/\rho}$. To sample V from ditribution p, we compute $V = \mathbf{G}^{\frac{1}{2}}U$, where $U \in \mathbb{R}^{r\times 3}$ is a random matrix, of which each element follows $\mathcal{N}(0,1)$, and $\mathbf{G}^{\frac{1}{2}} \approx \mathbf{Q}\Lambda^{\frac{1}{2}}$ is computed via the Nyström method.

We use $\lambda = 10$ or 50 and $\rho = 2$ in our experiments.

Network structure The network used in our experiment is a 5-layer point-wise multi-layer perceptron with a skip connection. The detailed structure is shown in Fig. 11.

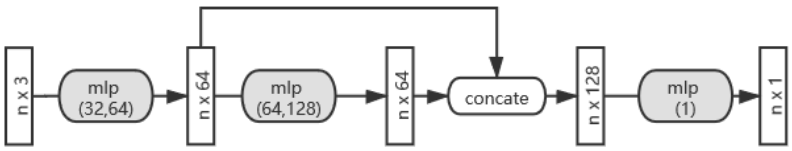


Figure 11: The structure of the network used in our experiments. The input is a matrix of shape $(n, 3)$ representing the coordinates of all points in the set, and the output is a matrix of shape $(n, 1)$ representing the potential of the corresponding points. $mlp(x)$ represents a multi-layer perceptron (mlp) with the size x. For example, $mlp(m, n)$ represents a mlp consisting of two layer, and the size of each layer is m and n. We use the ReLu activation function in all except the output layer.

Parameters We train the network $\boldsymbol{f}_{w,h}$ using the Adam optimizer (Kingma & Ba, 2014), and we train the transformation $\mathcal{T}_\theta$ using the RMSprop optimizer (Tieleman & Hinton, 2012). The learning rates of both optimizers are set to 10^{-4}.

For experiments in Sec. 5.3, the parameters as set as follows: For PWAN, we set $(\rho, \lambda, \sigma, T) = (2, 0.01, 0.1, 2000)$. For TPS-RPM, we set $T_finalfac = 500$, $frac = 1$, and $T_init = 1.5$. For GMM-REG, we set $sigma = 0.5, 0.2, 0.02$, $Lambda = .1, .02, .01$, $max_function_evals = 50, 50, 100$ and $level = 3$. For BCPD and CPD, we set $(\beta, \lambda, w) = (2.0, 2.0, 0.1)$.

For experiments in Sec. 5.4, the parameters as set as follows: For PWAN, we set $(\rho, \lambda, \sigma, T) = (1.0, 10^{-3}, 1.0, 3000)$. For BCPD and CPD, we set $(\beta, \lambda, w) = (3.0, 20.0, 0.1)$. We also test $w = 0.4$ for these methods, but the difference is not obvious.

For experiments in Sec. F.9, the parameters as set as follows: For PWAN, we set $(\rho, \lambda, \sigma, T) = (0.5, 5 \times 10^{-4}, 1.0, 3000)$. For BCPD and CPD, we set $(\beta, \lambda, w) = (3.0, 20.0, 0.1)$.

For experiments in Sec. F.10, the parameters as set as follows: We use m-PWAN. we set $(\rho, \lambda, \sigma, T) = (1.0, 1 \times 10^{-4}, 1.0, 3000)$. For BCPD and CPD, we set $(\beta, \lambda, w) = (0.3, 2.0, 0.1)$. We also tried $w = 0.6$ but the results are similar (not shown in our experiments).

We note the parameters for CPD and BCPD are suggested in Hirose (2021a).

Parameter selection The parameters m in m-PWAN and h in d-PWAN control the degree of alignment. Specifically, increasing m or h will lead to the registration of larger fraction of distributions. When $m = 0$ or $h = 0$, PWAN does not align any point (the gradients are 0 for all points); When $m = \min\{m_\alpha, m_\beta\}$ or $h = \infty$, PWAN seeks to align the largest fraction of points, *i.e.*, it becomes a DM method. Therefore, choosing overly large m or h may lead to the biased registration (like a DM method), while choosing overly small m or h may lead to the insufficient registration where too few points are aligned.

For uniform point sets, the parameters m and h can both be determined easily. For d-PWAN, if the averaged nearest distance between points is s, then h can be set near s. Because for the aligned point sets, the points that are farther from the overlapped region than distance s are likely to be outliers. For m-PWAN, one needs to estimate the overlap ratio ρ of one point set,*e.g.*, α, then m can be set as ρm_α. This is how we actually determined m and h in our experiments.

Refinement In our experiments, we find that a nearest-point-based refinement sometimes slightly improves the performance. Specifically, when the algorithm 1 is near convergence, we can optionally run a few steps of the nearest-point-based refinement to improve the performance slightly. Specifically, since algorithm 1 is near convergence, we can safely assume that the point sets are sufficiently aligned. Therefore, each point is highly likely to correspond to its nearest neighborhood in the other set within the mass or distance threshold. In other words, the correspondence matrix $\overline{\mathbf{P}} \in \mathbb{R}^{q \times r}$ can be written as

$$\overline{\mathbf{P}}_{i,j} = \begin{cases} 1 & if\ y_j = NN(x_i)\ and\ d(y_j, x_i) \leq h \\ 0 & else \end{cases} \tag{61}$$

for $\mathcal{L}_{D,h}$, and

$$\overline{\mathbf{P}}_{i,j} = \begin{cases} 1 & if\ y_j = NN(x_i)\ and\ x_i \in Nearest(m) \\ 0 & else \end{cases} \tag{62}$$

for $\mathcal{L}_{M,m}$, where $NN()$ is the nearest neighborhood function, and $Nearest(m)$ represent the nearest m points in X to Y. Thus we can slightly refine the result by switch the divergence term $\mathcal{L}$ to

$$\mathcal{L}(\alpha, \beta_\theta) = \sum_{i,j} \overline{\mathbf{P}}_{i,j} \boldsymbol{d}(x_i, \mathcal{T}_\theta(y_j))^2, \tag{63}$$

and update a few steps of the transformation $\mathcal{T}_\theta$ via gradient descent.

F.6 More details of Sec. 5.2

The KL divergence and the L_2 distance between point set X and Y are formally defined as

$$L_2(X,Y) = \sum_{\substack{x_i,x_j \in X \\ y_i,y_j \in Y}} \frac{1}{q^2}\phi(0|x_i - x_j, 2\sigma) + \frac{1}{r^2}\phi(0|y_i - y_j, 2\sigma) - \frac{2}{qr}\phi(0|x_i - y_j, 2\sigma), \tag{64}$$

$$KL(X,Y) = -\frac{1}{q}\sum_{y_j \in Y} \log\Big(\omega\frac{1}{q} + (1-\omega)\sum_{x_i \in X}\frac{1}{r}\phi(y_j|x_i,\sigma)\Big), \tag{65}$$

where $\phi(\cdot|u,\sigma)$ is the Gaussian distribution with mean u and variance σ. For simplicity, we set $\sigma = 1$ and $\omega = 0.2$ for KL and L_2.

F.7 More details of Sec. 5.3

We report the standard deviation (STD) of the MSE in experiments Sec. 5.3 in Tab. 4 and 5. It can be seen that our results generally have lower STDs than the baseline methods when the outlier ratio is high or when the overlap ratio is low (except for the bunny dataset in the first experiment where GMM-REG has the lowest STD). This implies that our method is more stable than the baseline methods under these situations.

Table 4: STD of the results of the first experiment.

(a) Bunny dataset

Outlier/Non-outlier ratio	0.2	0.4	0.6	0.8	1.0	1.2
BCPD	0.0897	0.1685	0.2647	6.131	6.0221	0.2099
CPD	**0.0043**	0.0363	0.0468	0.0607	0.0478	0.0602
GMM-REG	0.0055	**0.0068**	**0.0058**	**0.0064**	**0.0062**	**0.0089**
TPS-RPM	26.54	6.140	28.82	0.0275	0.0130	0.0952
PWAN	0.0149	0.0209	0.0369	0.0235	0.0402	0.0422

(b) Monkey dataset

Outlier/Non-outlier ratio	0.2	0.4	0.6	0.8	1.0	1.2
BCPD	0.2768	0.22	0.195	0.2776	0.4135	0.8966
CPD	0.0025	0.0175	0.022	**0.0338**	0.0373	0.0534
GMM-REG	**0.0025**	**0.0038**	**0.015**	0.0505	0.0941	0.1707
TPS-RPM	1977	8775	10207	116.1	2.318	0.551
PWAN	0.0267	0.0248	0.0317	0.0386	**0.0051**	**0.0488**

(c) Armadillo dataset

Outlier/Non-outlier ratio	0.2	0.4	0.6	0.8	1.0	1.2
BCPD	0.09	**0.0023**	0.2307	0.117	16.28	2.281
CPD	**0.0043**	0.0363	0.0468	0.0607	0.0478	0.0602
GMM-REG	0.0108	0.0114	0.0117	0.0245	0.0456	0.0588
TPS-RPM	1890	140.5	25.25	**0.0155**	1.124	0.0667
PWAN	0.0565	0.0532	**0.002**	0.0525	**0.0352**	**0.039**

We present some results of these two experiments with $N = 2000$ in Fig. 12 and Fig. 13, and the corresponding quantitative results are shown in Tab. 6 and Tab. 7. We do not show the results of TPS-RPM on the second experiment, as it generally fails to converge. As can be seen, PWAN successfully registers the point sets in all cases, while all baseline methods bias toward to the noise points or to the non-overlapped region when outlier ratio is high, except for TPS-RPM which shows strong robustness against noise points comparable with PWAN in the first example.

To obtain a sense of the learned potential network $\boldsymbol{f}_{w,h}$, we visualize the potential at the end of the registration process. The results are shown in Fig. 14 and Fig. 15. We represent the value of potential and the gradient norm at each point by colors, where brighter color indicates higher value. As can be seen in Fig. 15(c) and 14(c), the network assigns higher values to the points in α while assigning

Table 5: STD of the results of the first experiment.

(a) Bunny dataset

Overlap ratio	1.0	0.94	0.88	0.82	0.75	0.66	0.57
BCPD	0.0051	0.0068	0.0064	**0.0114**	0.0186	0.0319	**0.0472**
CPD	**0.001**	0.0053	0.0196	0.0619	0.0905	0.1905	0.2771
GMM-REG	0.0011	**0.0019**	**0.0031**	0.028	0.0792	0.1619	0.2376
TPS-RPM	1.9401	368.69	183129	7329	1317	1778	86995
PWAN-m	0.0342	0.0411	0.0264	0.0593	0.0489	0.0488	0.0576
PWAN-d	0.0457	0.0403	0.0455	0.0539	**0.043**	**0.0444**	0.0533

(b) Monkey dataset

Overlap ratio	1.0	0.94	0.88	0.82	0.75	0.66	0.57
BCPD	0.0042	0.006	**0.0069**	**0.0196**	0.0451	0.0747	0.1138
CPD	**0.0004**	0.0016	0.0132	0.0846	0.1403	0.1864	0.2243
GMM-REG	0.0005	**0.0012**	0.0155	0.0547	0.0881	0.149	0.2355
TPS-RPM	29319	265616	13233	2883	3122	5901	535
PWAN-m	0.026	0.0365	0.0313	0.0291	0.0536	0.0599	**0.045**
PWAN-d	0.015	0.0428	0.0437	0.0315	**0.0441**	**0.0415**	0.0466

(c) Armadillo dataset

Overlap ratio	1.0	0.94	0.88	0.82	0.75	0.66	0.57
BCPD	0.0025	0.0041	**0.006**	**0.0145**	**0.0287**	0.0574	0.084
CPD	**0.0003**	0.003	0.03	0.10	0.1617	0.3420	0.4358
GMM-REG	0.0015	**0.0023**	0.0411	0.0603	0.1517	0.2679	0.5094
TPS-RPM	375.2	24.81	22031	44.42	461.8	248.2	86890
PWAN-m	0.0224	0.0367	0.0363	0.0419	0.0482	0.043	**0.0501**
PWAN-d	0.0406	0.0373	0.041	0.0244	0.0487	**0.0224**	0.0691

lower values to the points in β_θ. In addition, as shown in Fig. 15(d) and 14(d), most of outliers, including noise points and the points in non-overlapping region, have low gradients, *i.e.*, they are successfully discarded by the network.

Table 6: Quantitative comparison of the registration results shown in Fig. 12 using MSE.

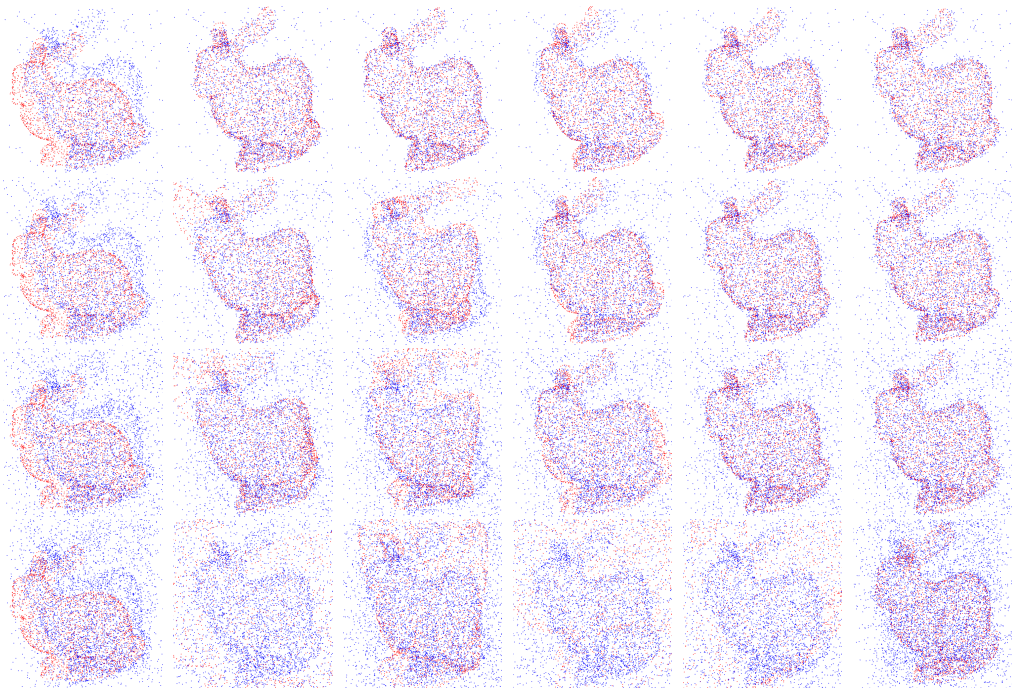
Outlier/Non-outlier ratio	BCPD	CPD	GMM-REG	TPS-RPM	PWAN
0.2	0.011	**0.0015**	0.018	0.0032	0.0031
0.6	0.075	0.083	0.022	**0.0026**	0.0030
1.2	0.12	0.21	0.042	0.0033	**0.0030**
2.0	1.24	0.289	3.76	2.354	**0.004**

Table 7: Quantitative comparison of the registration results shown in Fig. 13 using MSE.

Overlap ratio	BCPD	CPD	GMM-REG	d-PWAN	m-PWAN
0.57	0.18	0.92	0.32	0.017	**0.015**
0.75	0.028	0.45	0.043	**0.0044**	0.0090
1	**0.00072**	0.0025	0.0078	0.0037	0.0038

F.8 More details of Sec. 5.4

We provide more details regarding the computation time of our method. We first sample $q = r$ points from the bunny shape, where $q = 10^5$ or 7×10^5. We then run both the 1 GPU version and the 2 GPU version of PWAN 100 steps, and report their computation time in Fig. 16, where q-MGPU represents registering q points using M GPU. As can be seen, the majority of time is spent on updating the network, and the parallel training can effectively reduce the computation time for updating the network. In addition, the gain of speed increase as the number of points, *i.e.*, we get $1.3\times$ speedup when $q = 10^5$, while the speedup is $1.9\times$ when $q = 7 \times 10^5$, which is close to the theoretical speedup value 2.



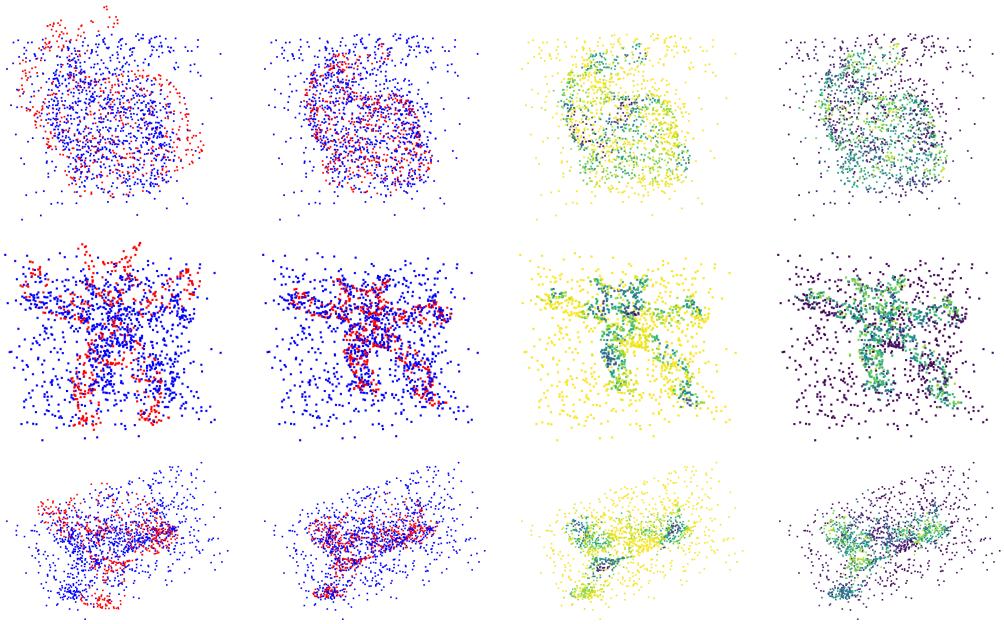
(a) Initial sets (b) BCPD (c) CPD (d) GMM-REG (e) TPS-RPM (f) PWAN

Figure 12: An example of registering noisy point sets. The outlier/non-outlier ratios of the point sets shown here are 0.2 (1st row), 0.6 (2nd row), 1.2 (3rd row) and 2.0 (4th row).

(a) Initial sets (b) BCPD (c) CPD (d) GMM-REG (e) d-PWAN (f) m-PWAN

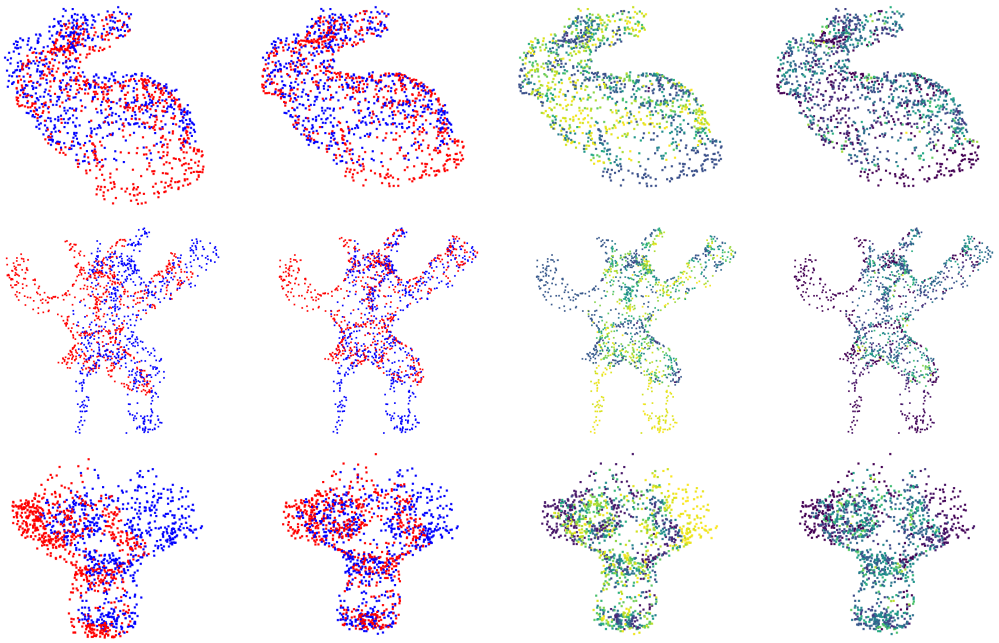
Figure 13: An example of registering partially overlapped point sets. The overlap ratio of the point sets shown here are 0.57 (1st row), 0.75 (2nd row) and 1 (3rd row).

Finally, we evaluate PWAN on large scale point sets. We generate noisy and partially overlapped armadillo datasets as stated in Sec. 5.3. We compare PWAN with BCPD and CPD, because they are the only baseline methods that are scalable in this experiment. We present some registration results in



(a) Initial sets (b) Registered sets (c) Potential $\boldsymbol{f}$ (d) $|\nabla \boldsymbol{f}|$

Figure 14: Visualization of the learned potentials on point sets with extra noise points.



(a) Initial sets (b) Registered sets (c) Potential $\boldsymbol{f}$ (d) $|\nabla \boldsymbol{f}|$

Figure 15: Visualization of the learned potentials on partially overlapped point sets.

Fig. 17. As can be seen, our method can handle both cases successfully, while both CPD and BCPD bias toward outliers.

More training details of PWAN is shown in Fig. 18, Fig. 19 and Fig. 20. As can be seen in Fig. 18(a), 19(a) and 20(a), the source sets are matched to the reference sets smoothly. Fig. 18(b), 19(b) and 20(b) suggest that at the end of the registration process, most of outliers are correctly discarded.

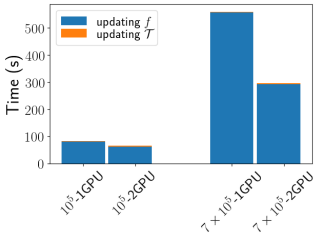
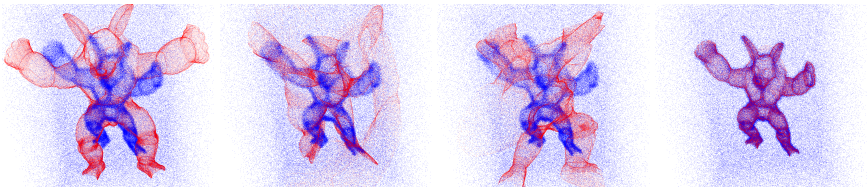
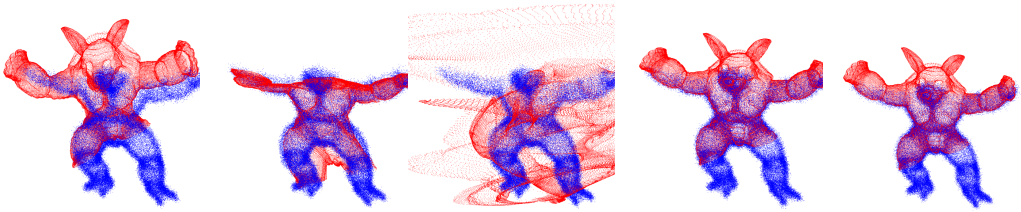


Figure 16: Computation time of our method.



Initial sets BCPD CPD PWAN

(a) An example of registering noisy point sets. The source and refernece sets contain 8×10^4 and 1.76×10^5 points respectively.



Initial sets BCPD CPD d-PWAN m-PWAN

(b) An example of registering partially overlapped point sets. The source and refernece sets both contain 7×10^4 points.

Figure 17: Examples of registering large scale point sets.

Table 8: Registration results of the space-time faces dataset.

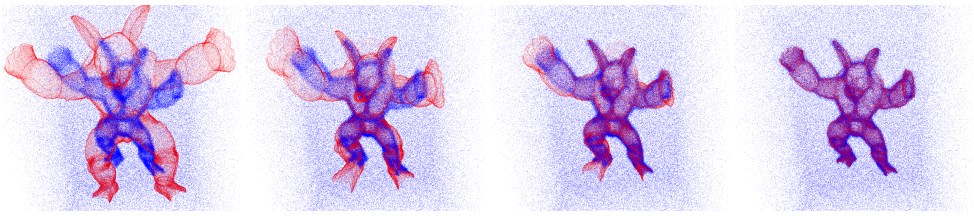
BCPD	CPD	PWAN
0.0017 ± 0.001	0.00049 ± 0.0002	0.00032 ± 0.000089

Besides, Fig. 18(c), 19(c) and 20(c) implies that the norm of gradient of the network is indeed controlled below 1, *i.e.*, it is indeed Lipschitz. In addition, both the training loss and the MSE decrease smoothly in the training process.

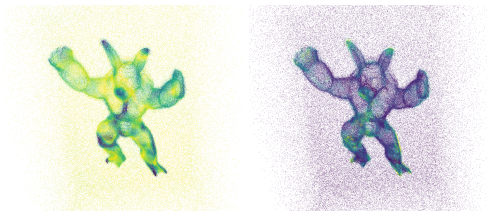
F.9 Experiment on the human face dataset

We evaluate our method in the space-time faces dataset (Zhang et al., 2008), which consists of a time series of point sets sampled from a real human face. Each face consists of $23,728$ points and the true correspondence between faces are known. We use the faces at time i and $i + 20$ as the source and the reference set, where $i = 1, ..., 20$. All point sets in this dataset are the same size and are completely overlapped.

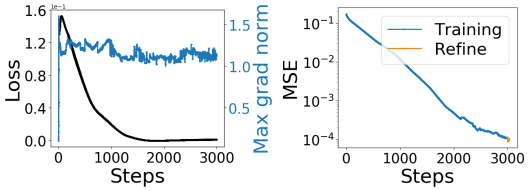
The registration results are shown in Tab. 8, where we can see PWAN outperforms both CPD and BCPD. We present the examples of the registration results in Fig. 21. As can be seen, PWAN successfully aligns the faces in different time points.



(a) Registration trajectory. The registration process proceedes from left to right.

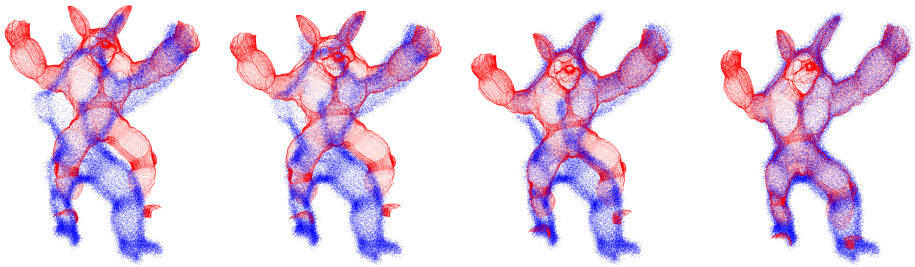


(b) Visualization of the potential $\boldsymbol{f}$ (left) and it gradient $|\nabla \boldsymbol{f}|$ (right) at the final step.

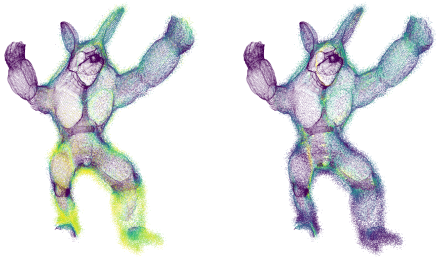


(c) Statistics of the training process.

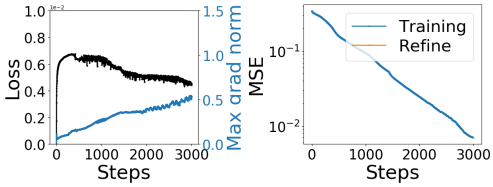
Figure 18: One example of registering the noisy "armadillo" datasets. The source and refernece sets contain 8×10^4 and 1.76×10^5 points respectively.



(a) Registration trajectory. The registration process proceedes from left to right.



(b) Visualization of the potential $\boldsymbol{f}$ (left) and it gradient $|\nabla \boldsymbol{f}|$ (right) at the final step.



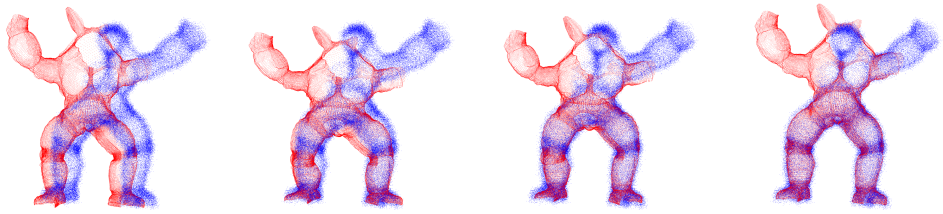
(c) Statistics of the training process.

Figure 19: One example of registering the partially overlapped "armadillo" datasets using d-PWAN. Each point set consists of 7×10^4 points.

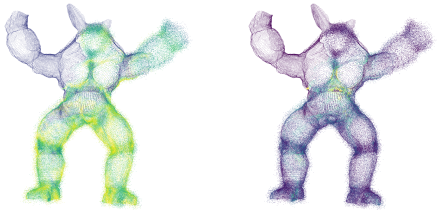
F.10 Experiment on the 3D human dataset

We evaluate our method on a challenging human shape dataset (DataSet), which is taken from a SHREC'19 track called "matching humans with different connectivity". This dataset consists of 44 shapes, and we manually select 3 pairs of shapes for our experiments. To generate a point set of each shape, we first sample 50000 random points from the surface of the 3D mesh, and then apply voxel grid filtering to down-sample the point set to less than 10000 points. The description for the selected point sets is presented in Tab. 9

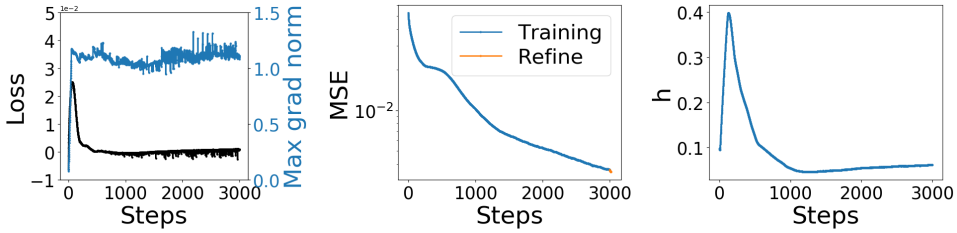
We conduct the following two experiments. In the first experiment, we evaluate our method on registering the complete point sets. In the second experiment, we evaluate our method on the more



(a) Registration trajectory. The registration process proceedes from left to right.



(b) Visualization of the potential $\boldsymbol{f}$ (left) and it gradient $|\nabla \boldsymbol{f}|$ (right) at the final step.



(c) Statistics of the training process.

Figure 20: One example of registering the partially overlapped "armadillo" dataset using m-PWAN. Each point set consists of 7×10^4 points.

Table 9: Point sets used for registration. no.m represents the m-th shape in the dataset (DataSet).

	(no.1, no.42)	(no.18, no.19)	(no.30, no.31)
Size	(5575, 5793)	(6090, 6175)	(6895, 6792)
Description	same pose different person	different pose same person	different pose different person

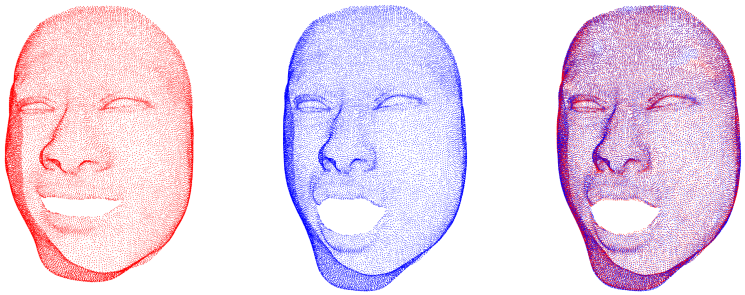
challenging partial matching problem, where we generate incomplete point sets by manually cropping a fraction of the no.30 and no.31 point sets. We consider 3 types of partial matching problem, *i.e.*, match incomplete set to complete set, match complete set to incomplete set and match incomplete set to incomplete set. For both of these two experiments, we compare our method with CPD and BCPD, and we only present qualitative registration results, because we do not know the true correspondence between point sets.

The results of the first experiment is shown in Fig. 22. As can be seen, PWAN can handle both the local deformations (1-st row) and the articulated deformations (2-nd and 3-rd rows) well, and it produces good full-body registration results. In contrast, although CPD and BCPD can handle local deformations relatively well (1-st row), they have difficulties aligning point sets with large articulated deformations, as significant registration errors are observed near the limbs (2-nd and 3-rd rows).

The results of the second experiment is shown in Fig. 23. As can be seen, both CPD and BCPD fail in this experiment, as the non-overlapping points are seriously biased. For example, in the 3-rd row, they both wrongly match the left arm to the body, which causes highly unnatural artifacts. In contrast, the proposed PWAN can handle the partial matching problem well, since it successfully maintains the shape of non-overlapping regions, which contributions to the natural registration results.

Finally, we note that although the proposed PWAN generally produces reasonable full-body registration results, it has some difficulties handling the local details. For example, the hands in Fig. 22 and 23

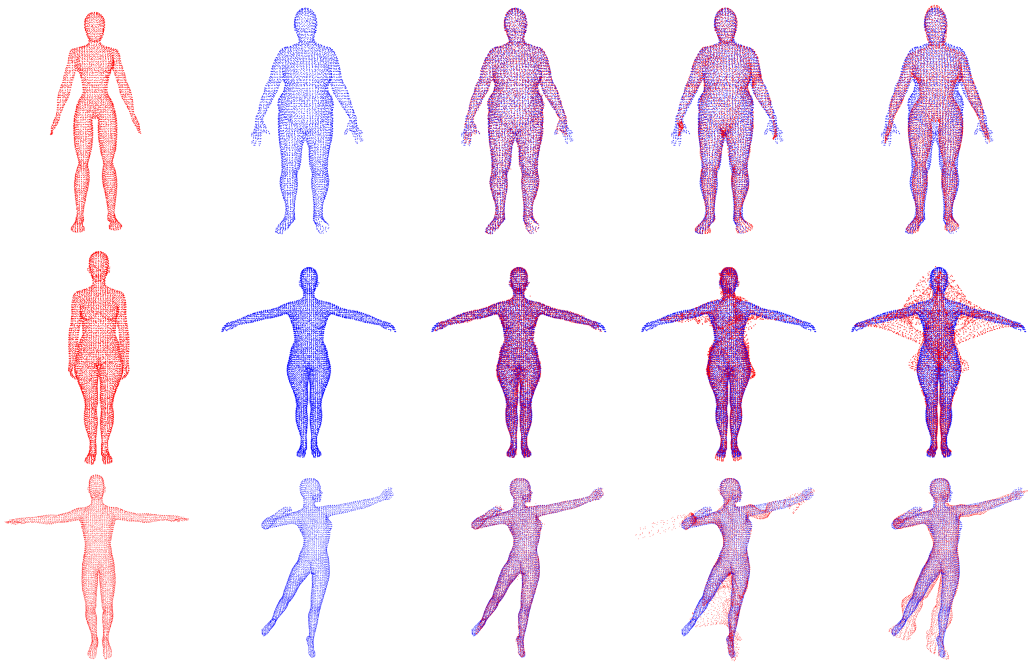
(a) The aligned point sets (right) is obtained by matching the 1-st frame (left) to the 21-st frame (middle)



(b) The aligned point sets (right) is obtained by matching the 19-st frame (left) to the 39-st frame (middle)

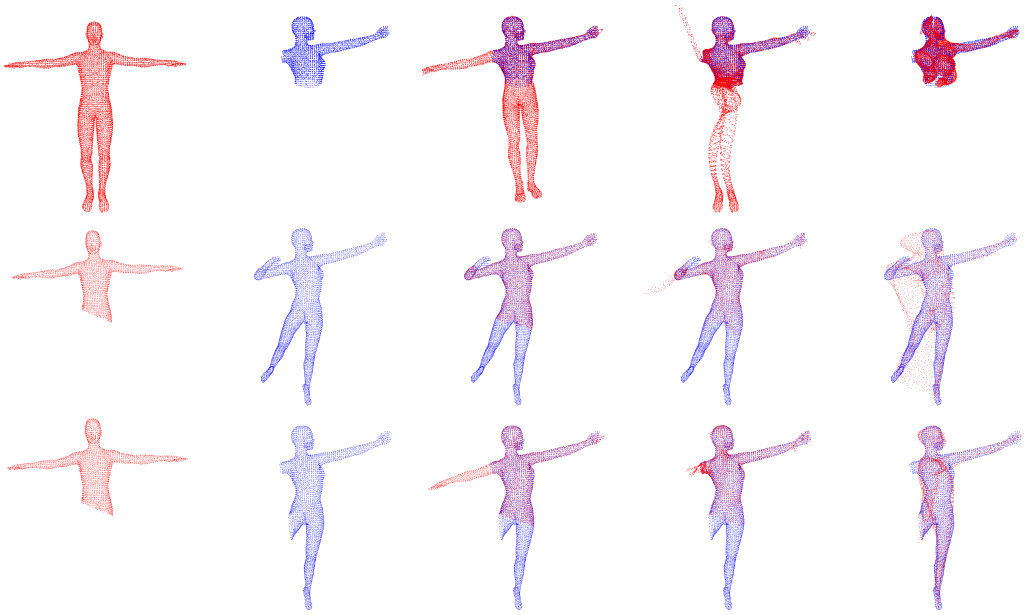
Figure 21: Examples of our registration results on the spatial-temporal faces dataset. Zoom in to see the details.

are generally not well aligned. This drawback might be alleviated by considering local constraints such as Ge et al. (2014) in the future. In addition, it is worth noticing that the aligned point sets in the second experiments are natural and do not exist in the original dataset. These results suggests the potential of our method in other practical tasks such as point set completion and point sets merging.



(a) Source set (b) Reference set (c) PWAN (Ours) (d) BCPD (e) CPD

Figure 22: The results of registering complete point sets no.1 to no.42 (1-st row), no.18 to no.19 (2-nd row), and no.30 to no.31 (3-rd row). Our results are compared against BCPD (Hirose, 2021a) and CPD (Myronenko & Song, 2010). Zoom in to see the details.



(a) Source set (b) Reference set (c) PWAN (Ours) (d) BCPD (e) CPD

Figure 23: Registering incomplete point sets no.30 to no.31. We present the results of the complete to incomplete (1-st row), incomplete to complete (2-nd row) and incomplete to incomplete (3-rd row) registration. Our results are compared against BCPD (Hirose, 2021a) and CPD (Myronenko & Song, 2010). Zoom in to see the details.