Scalable Knothe–Rosenblatt-like Heuristic Transportation Plans for Imaging Problems

Gennaro Auricchio
University of Padua
Padua, Italy
gennaro.auricchio@unipd.it

Min Lin
Xiamen University of Technology
Xiamen, China
minlin@stu.xmut.edu.cn

Lingxuan Zhou
Xiamen University of Technology
Xiamen, China
lingxuanzhou@stu.xmut.edu.cn

Zhaori Guo*
Xiamen University of Technology
Xiamen, China
zhaoriguo@xmut.edu.cn

Zhongqi Cai*
Xiamen University of Technology
Xiamen, China
zhongqicai@xmut.edu.cn

ABSTRACT

In this paper, we introduce a novel formalism for computing the Wasserstein Distance between any pair of probability distributions, μ and ν. Standard approaches require solving a matching problem between two discrete distributions, which becomes computationally expensive as the dimensionality increases. To address this challenge, we propose a new family of heuristic transportation plans that extend the classic Knothe–Rosenblatt transport plan. Each heuristic plan is associated with a method for combining the two original measures into an *intermediate measure*, significantly reducing the number of variables required to characterise any transportation plan. Specifically, if the probability measures μ and ν have supports consisting of N and M points, respectively, our approach reduces the number of variables from $N \times M$ to $\min\{N, M\}$. We demonstrate that our method is particularly well-suited for defining a neural network to solve the optimal transport problem and validate our model through extensive numerical experiments.

KEYWORDS

Wasserstein Distance; Optimal Transport; Knothe–Rosenblatt transportation plan; Image Generation

ACM Reference Format:
Gennaro Auricchio, Min Lin, Lingxuan Zhou, Zhaori Guo*, and Zhongqi Cai*. 2026. Scalable Knothe–Rosenblatt-like Heuristic Transportation Plans for Imaging Problems. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/CEMK9641

1 INTRODUCTION

In its modern formulation, the Optimal Transport (OT) problem is postulated as a Linear Programming (LP) problem which reads as

$$\min_{\pi \in \Pi(\mu,\nu)} \left(\sum_{\mathbf{i},\mathbf{j}} ||\mathbf{i} - \mathbf{j}||_p^p \pi_{\mathbf{i},\mathbf{j}} \right)^{\frac{1}{p}}, \tag{1}$$

where $\Pi(\mu,\nu) := \left\{ \pi \in \mathcal{P}([n]^2 \times [n]^2) : \sum_{\mathbf{j}} \pi_{\mathbf{i},\mathbf{j}} = \mu_{\mathbf{i}}, \sum_{\mathbf{i}} \pi_{\mathbf{i},\mathbf{j}} = \nu_{\mathbf{j}} \right\}$. From an applied perspective, a key advantage of the Optimal Transportation problem lies in the fact that (1) defines a distance over the space of probability measures known as the p-th Wasserstein Distance W_p [51]. The W_p distance has become a powerful tool in computer science, enabling applications such as quantifying the similarity between images [17, 40, 44], assessing model convergence [8, 9], and training generative adversarial networks (GANs) [26, 41, 52]. A major limitation of the Wasserstein distance is its high computational cost, since evaluating it requires solving a Linear Programming problem. The growing use of Wasserstein distances has therefore spurred significant interest in developing efficient algorithms to compute this distance. The most widely used methods include: (i) Sinkhorn's algorithm [3, 17, 49], which solves a regularised version of the optimal transport problem; (ii) Linear Programming-based algorithms [6, 13, 34], which solve the optimal transport problem as an uncapacitated minimum cost flow problem [24, 38]; and (iii) Linear Programming-based approximations [7, 40], which aim to reduce the computational complexity of LP-based algorithms by simplifying the problem structure.

A fourth and less explored method is to define suitable heuristics that, albeit being sub-optimal, possess other relevant properties, such as being easy to compute or having triangular gradient. In Optimal Transport, a heuristic is a function $\mathcal{S}$ that maps any couple of probability distributions (μ, ν) to an element of $\Pi(\mu, \nu)$, *i.e.* a feasible transportation plan between μ and ν. The most famous example of such heuristics is the Knothe–Rosenblatt transportation plan which was introduced by Knothe [32] and Rosenblatt [43] independently. Unlike cost-optimal OT solutions, the KR plan prioritises computational efficiency, making it particularly attractive for large-scale applications. Indeed, the Knothe–Rosenblatt (KR) rearrangement is constructed by sequentially matching marginals and conditional marginals along each coordinate, allowing to bypass the solution of an LP problem [46]. This recursive coordinate-wise reordering yields a transport map with a triangular structure that can be computed in $O(n^2)$ time. Owing to these properties, the Knothe–Rosenblatt heuristic is an appealing cheap-to-compute alternative to the optimal transportation plan. In this paper, we introduce a novel class of heuristic transportation plans that generalise the Knothe–Rosenblatt heuristic, offering a variety of different tools without altering its computational efficiency.

Correspondence to Zhaori Guo <zhaoriguo@xmut.edu.cn> and Zhongqi Cai <zhongqicai@xmut.edu.cn>.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/CEMK9641

1.1 Related Works

Over the past few decades, the Optimal Transport (OT) problem [31, 36] has found numerous applications across diverse fields, including Machine Learning [4, 18, 23, 47], Computer Vision [40, 45], Partial Differential Equations [15, 16, 35], and Algorithmic Game Theory [9, 19]. The increasing relevance of these distances in Computer Science has driven a growing demand for efficient solutions or sensible approximations to the minimisation problem that defines them. Currently, several approaches exist for solving the minimisation problem that defines the Wasserstein Distance. The most widely used methods include: (i) *Sinkhorn-based algorithms* [3, 17, 49], which compute a regularised version of the optimal transport problem, (ii) *Linear Programming-based algorithms* [6, 13, 34], which approximate or exactly solve the optimal transport problem by formulating it as an equivalent uncapacitated minimum-cost flow problem [24, 38], and (iii) *Linear Programming-based approximations*, which reduce the computational complexity of LP-based methods by simplifying the problem structure, as in [7, 40]. These three approaches exhibit distinct characteristics that make each of them relevant in different contexts. For instance, the Sinkhorn algorithm is the fastest among the three; however, its accuracy depends on selecting an appropriate regularisation parameter [17]. In contrast, Linear Programming (LP) methods yield the exact optimal solution and require no parameter tuning, but their computational cost is often prohibitive unless the problem possesses certain separability properties [6, 34]. Finally, LP-based simplifications offer a compromise between these two extremes by solving a simpler LP problem, usually relying on a truncation parameter [7, 40] or the number of directions used in the approximation [13]. In this paper, we consider a heuristic-based approach that extends the Knothe–Rosenblatt rearrangement plan [32, 43]. Despite being a heuristic solution, the Knothe–Rosenblatt transportation plan has been fruitful in several applied domains such as income analysis [28], density estimation [21, 50], stochastic process analysis [10], and machine learning [2, 11]. Furthermore, these constructions have found wide usage in normalising flows [33, 39, 48], specifically in autoregressive normalising maps [30]. All these works showcase the power of having a suitable and easy-to-compute heuristic.

Lastly, we notice that OT problems and related matching problems are well established in the study of multi-agent systems, where they arise in facility location, social optimisation problems (SOPS), and mechanism design (see, e.g., [9, 19, 25]). Our theoretical contribution is aligned with this line of work, as it introduces OT-based tools that can be used to characterise and compute an optimal allocation rule. In parallel, neural network models have recently attracted growing interest in the multi-agent systems community as auxiliary components for decision-making, particularly in learning-augmented and prediction-based mechanism design (e.g., [1, 12, 53]).

1.2 Our Contribution

In this paper, we propose a novel class of heuristic transportation plans that extends and generalises the well-known Knothe–Rosenblatt rearrangement [32, 43]. Each element of our family is characterised by a composition rule, i.e. a function $\mathcal{R}$ that maps every couple of two probability distributions, namely μ and ν, into a suitable third probability distribution, namely ζ. Indeed, we demonstrate that every composition rule generates a unique function that maps any pair of probability distributions to a feasible transportation plan between the two input probability distributions. We first show that our formalism encompass all the relevant already-known transportation plans, such as the Knothe–Rosenblatt rearrangement and the optimal transportation plan, and generates new heuristic transportation plans. We corroborate our study by showcasing the main theoretical properties of our approach. In particular, we show that every composition rule is associated with the solution to a suitable optimal transportation problem and provide relevant error bounds. To support our theoretical study, we leverage our results to design a Neural Network model capable of approximating the optimal transportation plan. This model serves as an oracle for a computationally demanding problem, enabling a range of applications, including image clustering and classification. We illustrate its effectiveness through the task of comparing medical images. For clarity of exposition, we focus on probability measures supported on a two-dimensional space, although the methodology naturally extends to higher-dimensional settings.

Algorithm 1 1dOPT Algorithm [14]

Require: Measures (x, α, n) and (y, β, m), cost C
Ensure: Primal solution π

```
Set π ← 0,
Set a, b, i, j ← α_1, β_1, 1, 1
while i < n or j < m do
    if (a > b and i < n) or (j = m) then
        π_{i,j}, b, a ← a, b − a, α_{i+1}
        i ← i + 1
    else if (a > b and j < m) or (i = n) then
        π_{i,j}, a, b ← b, a − b, β_{j+1}
        j ← j + 1
    end if
end while
return π
```

2 PRELIMINARIES

For simplicity, we restrict our discussion to the case where probability measures are supported on two $n \times n$ grids; the extension to higher-dimensional settings is reported in the Appendix.

Under these assumptions, a probability measure is defined as $\mu := \sum_{i_1,i_2\in[n]} \mu_{\mathbf{i}}\, \delta_{x_{\mathbf{i}}}$, where $\{x_{\mathbf{i}}\}_{\mathbf{i}\in[n]^2} = \left\{\left(\frac{i_1}{n}, \frac{i_2}{n}\right)\right\}_{i_1,i_2\in[n]}$ and $\mu_{\mathbf{i}} \geq 0$ are such that $\sum_{\mathbf{i}\in[n]^2} \mu_{\mathbf{i}} = 1$. We denote with $\mathcal{P}([n]^2)$ the set of probability measures supported over the two-dimensional grid. The first and second marginals of μ are then defined as

$$\mu^{(1)}_{i_1} = \sum_{i_2=1}^{n} \mu_{i_1,i_2} \qquad \text{and} \qquad \mu^{(2)}_{i_2} = \sum_{i_1=1}^{n} \mu_{i_1,i_2}$$

respectively, so that the conditional measure of μ with respect to the first and second components are given by $(\mu_{|i_1})_{i_2} = \mu_{i_1,i_2}/\mu^{(1)}_{i_1}$ and $(\mu_{|i_2})_{i_1} = \mu_{i_1,i_2}/\mu^{(2)}_{i_2}$, respectively. With a slight abuse of notation, we write $\mu = \mu_{|\,i_j} \otimes \mu^{(j)}$. Finally, for any $p \geq 1$, we define the p-th power of the ℓ_p-norm of a vector $\mathbf{x} = (x_1, x_2)$ as $\|\mathbf{x}\|_p^p = |x_1|^p + |x_2|^p$.

2.1 The Wasserstein Distances

We now review the main notions about optimal transport. First, we introduce the notion of transportation plan between two measures.

Definition 1. *Let μ and ν be two probability measures over the two-dimensional regular grid $[n]^2$. A probability measure $\pi \in \mathcal{P}([n]^2 \times [n]^2)$ is said to be a transportation plan between μ and ν if*

$$\mu_{\boldsymbol{i}} = \sum_{\boldsymbol{j} \in [n]^2} \pi_{\boldsymbol{i},\boldsymbol{j}} \qquad \text{and} \qquad \nu_{\boldsymbol{j}} = \sum_{\boldsymbol{i} \in [n]^2} \pi_{\boldsymbol{i},\boldsymbol{j}}. \tag{2}$$

We denote with $\Pi(\mu, \nu)$ the set of π satisfying conditions (2).

It is straightforward to verify that $\Pi(\mu, \nu)$ is always non-empty. Indeed, the independent transportation plan $\pi = \mu \otimes \nu$, defined by $\pi_{\mathbf{i},\mathbf{j}} = \mu_{\mathbf{i}} \nu_{\mathbf{j}}$, constitutes a valid transportation plan between μ and ν. Consequently, the minimisation problem that defines the Wasserstein distance in (1) is well-posed [51].

Definition 2. *Given $p \geq 1$, let $\mu, \nu \in \mathcal{P}([n]^2)$. Denoted with $\mathbb{T}_p : \mathcal{P}([n]^4) \to \mathbb{R}$ the transportation cost functional, that is $\mathbb{T}_p(\pi) = \sum_{\boldsymbol{i},\boldsymbol{j}} ||\boldsymbol{i} - \boldsymbol{i}||_p^p \pi_{\boldsymbol{i},\boldsymbol{j}}$, the Wasserstein Distance between μ and ν, namely $W_p(\mu, \nu)$, is defined as $\big(W_p(\mu, \nu)\big)^p := \min_{\pi \in \Pi(\mu,\nu)} \mathbb{T}_p(\pi)$.*

When $d = 1$, $W_p(\mu, \nu)$ is expressed through a closed formula. In particular, it is possible to compute the optimal transportation plan via an algorithm (1dOPT) that we report in Algorithm 1, as shown in [14]. Notice that 1dOPT runs in $O(n)$ time.

Although for $d \geq 2$ no closed-form solution exists for the optimal transportation plan, it is still possible to decompose the plan into two one-dimensional transportation plans [5].

Theorem 1 (Theorem 2, [5]). *Given $p \geq 1$, let μ, ν be two probability distributions and let $\mathcal{I}(\mu, \nu)$ be defined as*

$$\mathcal{I}(\mu, \nu) := \left\{ \zeta \in \mathcal{P}([n]^2) \;\; s.t. \;\; \sum_a \zeta_{a,b} = \sum_a \mu_{a,b};\; \sum_b \zeta_{a,b} = \sum_b \nu_{a,b} \right\}.$$

Then, there exists $\zeta \in \mathcal{I}(\mu, \nu)$ for which the following identity holds

$$W_p^p(\mu, \nu) = \sum_{i_2} W_p^p(\mu_{|i_2}, \zeta_{|i_2}) \mu_{i_2}^{(2)} + \sum_{j_1} W_p^p(\zeta_{|j_1}, \nu_{|j_1}) \nu_{j_1}^{(1)}. \tag{3}$$

2.2 The Knothe–Rosenblatt Rearrangement

Given $\mu, \nu \in \mathcal{P}([n]^2)$, the *Knothe–Rosenblatt (KR) transport plan* (or Knothe–Rosenblatt rearrangement) provides a way to map μ to ν by iteratively matching their marginals. In the discrete setting, the construction proceeds as follows:

(A) **Compute the first marginals.** Given two probability measures μ and ν supported on the same two-dimensional grid $[n]^2$, we compute their first marginals $\mu^{(1)}$ and $\nu^{(1)}$.

(B) **Match the first marginals.** Using Algorithm 1, we compute the optimal transportation plan $\gamma^{(1)}$ between $\mu^{(1)}$ and $\nu^{(1)}$.

(C) **Adjust the source distribution.** We then use $\gamma^{(1)}$ to transform μ into a distribution whose first marginal matches $\nu^{(1)}$. Specifically, we introduce the following transportation plan

$$\tilde{\gamma}^{(1)}_{i_1,i_2,j_1,j_2} = \begin{cases} (\mu_{|i_1})_{i_2} \, \gamma^{(1)}_{i_1,j_1}, & \text{if } i_2 = j_2, \\ 0, & \text{otherwise.} \end{cases}$$

(D) **Compute the intermediate distribution.** We define

$$\eta_{j_1,j_2} = \sum_{i_1,i_2} \tilde{\gamma}^{(1)}_{i_1,i_2,j_1,j_2},$$

which represents the distribution obtained after the rearrangement step described in point (C).

(E) **Match the conditional distributions.** For each j_1, we use Algorithm 1 to compute the optimal transport plan $\gamma^{(j_1)}$ between the conditional distributions $\nu_{|j_1}$ and $\eta_{|j_1}$ and set

$$\tilde{\gamma}^{(2)}_{i_1,i_2,j_1,j_2} = \begin{cases} \gamma^{(j_1)}_{i_2,j_2} \, \nu^{(1)}_{j_1}, & \text{if } i_1 = j_1, \\ 0, & \text{otherwise.} \end{cases}$$

(F) **Combine the steps.** The final KR transport plan is obtained by composing the two stages:

$$(\pi_{KR})_{i_1,i_2,j_1,j_2} = \frac{\tilde{\gamma}^{(1)}_{i_1,i_2,j_1,j_2} \, \tilde{\gamma}^{(2)}_{i_1,i_2,j_1,j_2}}{\eta_{j_1,j_2}}.$$

The resulting plan π_{KR} has a *triangular structure*: the assignment of the j-th coordinate depends only on the first j coordinates. Figure 2 illustrates the KR rearrangement applied to two discrete measures. For a detailed treatment of the Knothe–Rosenblatt rearrangement in the continuous setting, we refer the reader to [51].

3 THE NEW FAMILY OF HEURISTIC TRANSPORTATION PLANS

In this section, we propose a new family of heuristic transportation plans between measures over the euclidean space that generalises the Knothe–Rosenblatt heuristic. Our approach is based on the notion of composition rule, which extends the techniques used to build the measure η in step (D) of the Knothe–Rosenblatt procedure.

3.1 The Composition Rule

The key notion behind our family of heuristics is the *composition rule*, which is a function that given in input two probability measures returns a third probability measure.

Definition 3. *Given $n \in \mathbb{N}$, a composition rule for $\mathcal{P}([n]^2)$ is a map $\mathcal{R} : \mathcal{P}([n]^2) \times \mathcal{P}([n]^2) \to \mathcal{P}([n]^2)$ such that $\mathcal{R}(\mu, \nu) \in \mathcal{I}(\mu, \nu)$, for every input measures μ and ν.*

Given two probability measures μ and ν and denoted by $\zeta \in \mathcal{I}(\mu, \nu)$ the output of a composition rule $\mathcal{R}$, we can mimic the KR rearrangement routine to build a transportation plan between μ and ν in $O(n^2)$ time. For any i_2, let us now consider the two conditional probability distributions $\mu_{|i_2}$ and $\zeta_{|i_2}$. Using 1dOPT (see Algorithm 1) we retrieve an optimal transportation plan between $\mu_{|i_2}$ and $\zeta_{|i_2}$, namely $\gamma_{|i_2}$, in $O(n)$ time. We thus retrieve a family of optimal transportation plans for every i_2, namely $\{\gamma^{(1)}_{|i_2}\}_{i_2 \in [n]}$, and define

$$\gamma^{(1)}_{i_1,i_2,j_1} = \gamma^{(1)}_{|i_2} \otimes \mu^{(2)}_{i_2} = \begin{cases} (\gamma^{(1)}_{|i_2})_{i_1,j_1} \cdot \mu^{(2)}_{i_2} & \text{if } \mu^{(2)}_{i_2} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

By definition, $\gamma^{(1)}$ is a probability measure that depends on (i_1, i_2, j_1) such that (i) the marginal on the first two entries is equal to μ, (ii) the marginal on the last two entries is equal to ζ, and (iii) the conditional law of $\gamma^{(1)}$ with respect to i_2 is the optimal transportation plan between $\mu_{|i_2}$ and $\zeta_{|i_2}$. We then apply the same construction to ζ and ν: for every j_1 we compute the conditional law of ν and ζ with respect to j_1 and the optimal transportation plan between $\zeta_{|j_1}$ and $\nu_{|j_1}$. In particular, we retrieve a probability measure $\gamma^{(2)} := \{\gamma^{(2)}_{i_2,j_1,j_2}\}$ such that the marginal on the first two

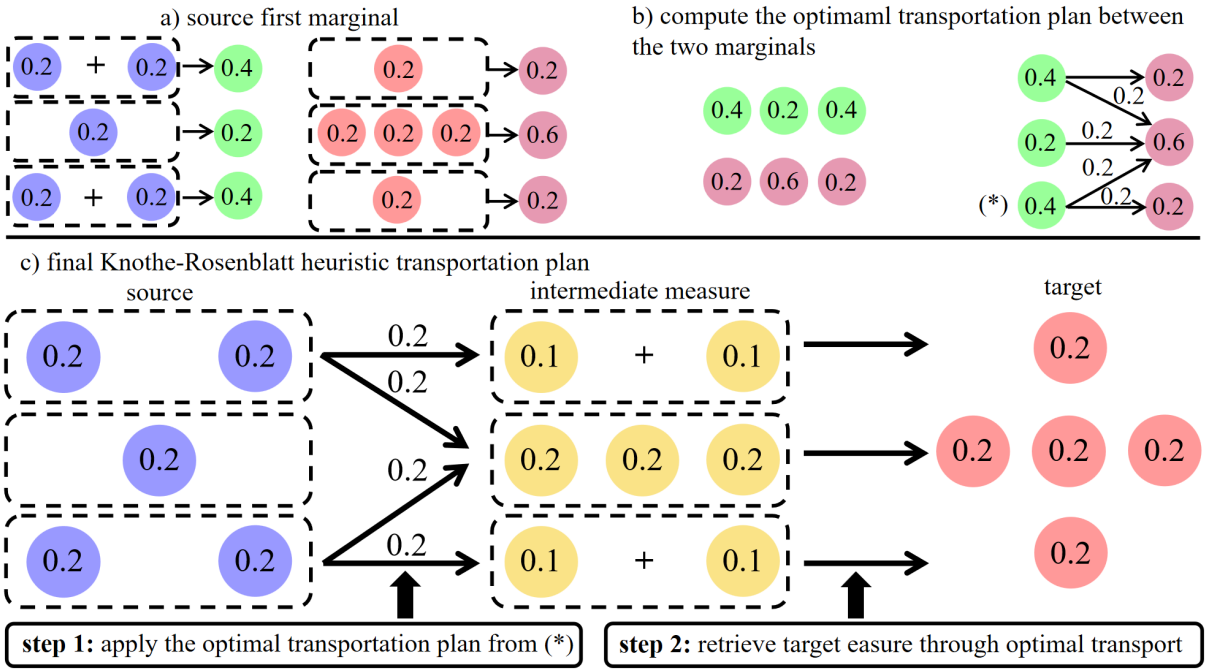


Figure 1: The evolution of the source distribution into the target distribution under the action of the Knothe–Rosenblatt rearrangement. Blue dots represent the source distribution on a 3×3 grid, while the red dots the target distribution supported over the same 3×3 grid. The yellow points represent the intermediate measure built at step (D). Since the marginals $\mu^{(1)}$ and $\nu^{(1)}$ are $(0.4, 0.2, 0.4)$ and $(0.2, 0.6, 0.2)$, respectively, the optimal transportation plan between these two measures moves 0.2 from the first coordinate into the second one and 0.2 form the third coordinate into the second one. By applying this transformation to the complete source distribution μ, we retrieve the intermediate measure from step (D).

entries is equal to ζ, the marginal on the last two entries is equal to ν, and the conditional law of $\gamma^{(2)}$ with respect to j_1 is an optimal transportation between $\zeta_{|j_1}$ and $\nu_{|j_1}$.

To conclude, for any given μ, ν, and $\zeta \in \mathcal{I}(\mu, \nu)$, we define π as

$$\pi(\mu, \nu; \zeta) := \frac{\gamma^{(1)} \otimes \gamma^{(2)}}{\zeta} = \begin{cases} \frac{\gamma^{(1)}_{i_1,i_2,j_1} \cdot \gamma^{(2)}_{i_2,j_1,j_2}}{\zeta_{j_1,i_2}} & \text{if } \zeta_{j_1,i_2} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

We summarise this procedure in Algorithm 2 and characterise its time complexity.

Proposition 1. *Algorithm 2 takes $O(n^2)$ operations to complete.*

To conclude this section, we show that Algorithm 2 allows us to associate any composition rule $\mathcal{R}$ to a map that sends any couple of probability measures into an element of $\Pi_{\mathcal{R}(\mu,\nu)}(\mu, \nu)$.

Theorem 2. *Given $\mu, \nu \in \mathcal{P}([n]^2)$ and a composition rule $\mathcal{R}$, the map $\mathcal{S}_{\mathcal{R}} : \mathcal{P}([n]^2) \times \mathcal{P}([n]^2) \to \mathcal{P}([n]^2 \times [n]^2)$, defined as*

$$\mathcal{S}_{\mathcal{R}}(\mu, \nu) = \pi(\mu, \nu; \mathcal{R}(\mu, \nu)), \tag{6}$$

where $\pi(\mu, \nu; \mathcal{R}(\mu, \nu))$, which is defined as in (5), is such that $\mathcal{S}_{\mathcal{R}}(\mu, \nu) \in \Pi_{\mathcal{R}(\mu,\nu)}(\mu, \nu)$ for every couple of probability measures μ and ν.

Theorem 2 gives us a general rule to build a family of heuristic transportation plan given any composition rule $\mathcal{R}$. We name these heuristics *KR-like Transportation Heuristics*, since by selecting a suitable composition rule $\mathcal{R}$ it is possible to retrieve the Knothe–Rosenblatt heuristic transportation plan.

3.2 Three Examples of Composition Rules

Next, we showcase three relevant examples of composition rules.

3.2.1 The Knothe–Rosenblatt Composition Rule. We show that the Knothe–Rosenblatt heuristic naturally arises from our composition rule, thereby demonstrating that our approach generalises the structure of the Knothe–Rosenblatt transportation plan.

Definition 4. *We define the Knothe–Rosenblatt composition rule $\mathcal{R}_{KR} : \mathcal{P}([n]^2) \times \mathcal{P}([n]^2) \to \mathcal{P}([n]^2)$ as*

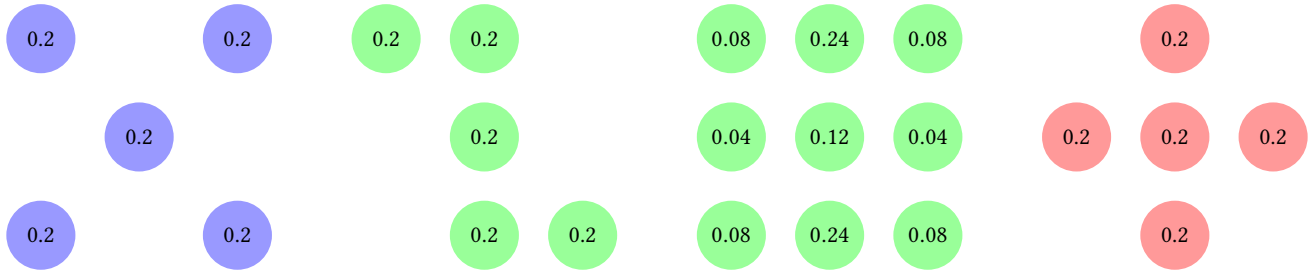
$$\mathcal{R}_{KR}(\mu, \nu) := \sum_{i_1} (\mu_{|i_1})_{i_2} \gamma_{i_1,j_1}, \tag{7}$$

where γ is the optimal transportation plan between $\mu^{(1)}$ and $\nu^{(1)}$.

Next, we show that $\mathcal{R}_{KR}$ is a composition rule and that the operator $\mathcal{S}_{\mathcal{R}_{KR}}$ is the Knothe-Rosenblatt transportation plan.

Theorem 3. *The function $\mathcal{R}_{KR}$ is a composition rule. Moreover, for every $\mu, \nu \in \mathcal{P}(\mathbb{R}^2)$, we have that $\mathcal{S}_{KR}(\mu, \nu) = \pi_{KR}$, where π_{KR} is the Knothe–Rosenblatt transportation plan.*

3.2.2 The Independent Composition Rule. Given two probability measures μ and ν, a classic transportation plan is the independent



(a) Source distribution μ (b) Optimal composition rule (c) Independent composition rule (d) Target distribution ν

Figure 2: Blue dots represent the source distribution on a 3×3 grid, while the red dots the target distribution supported over the same 3×3 grid. The green dots represent two possible outputs of a composition rule between the source and target measure. The intermediate measure on the left (subfigure b) is the output of an optimal composition rule. The intermediate measure on the right (subfigure c) is the intermediate measure returned by the independent composition rule. Notice that the outputs of both composition rules are such that their first marginal is $\nu^{(1)}$ and the second marginal is $\mu^{(2)}$.

Algorithm 2 KR-like Heuristic Transportation Plan Algorithm

Require: Measures $\mu, \nu, \zeta \in \mathcal{I}(\mu, \nu)$ and a cost C
Ensure: A transportation plan $\pi \in \Pi_\zeta(\mu, \nu)$

Set $\pi, f^{(1)}, f^{(2)} \leftarrow 0$,
for $i_2 \in [n]$ **do**:
  $\gamma^{(i_2)} \leftarrow 1dALG(\mu_{|i_2}, \zeta_{|i_2})$
  for $i_1, j_1 \in [n]$ **do**:
    $f^{(1)}_{i_1,i_2,j_1} \leftarrow \gamma^{(i_2)}_{i_1,j_1} \mu^{(2)}_{i_2}$
  end for
end for
for $j_1 \in [n]$ **do**:
  $\gamma^{(j_1)} \leftarrow 1dALG(\zeta_{|j_1}, \nu_{|j_1})$
  for $i_2, j_2 \in [n]$ **do**:
    $f^{(2)}_{i_2,j_1,j_2} \leftarrow \gamma^{(j_2)}_{i_2,j_2} \nu^{(1)}_{j_1}$
  end for
end for
for $i_1, i_2, j_1, j_2 \in [n]$ **do**
  $\pi_{i_1,i_2,j_1,j_2} \leftarrow \frac{f^{(1)}_{i_1,i_2,j_1} f^{(2)}_{i_2,j_1,j_2}}{\zeta_{i_2,j_1}}$ if $\zeta_{i_2,j_1} > 0$, $\pi_{i_1,i_2,j_1,j_2} \leftarrow 0$ otherwise.
end for
return π

one, defined as $\pi_{i_1,i_2,j_1,j_2} = \mu_{i_1,i_2} \otimes \nu_{j_1,j_2}$. Following the standard notation, we will denote the independent plan as $\pi = \mu \otimes \nu$. Through the use of the composition rules, we build a more efficient transportation plan between μ and ν.

Definition 5. *We denote with $\mathcal{R}_{ind}$ the independent composition rule, which is defined as $\mathcal{R}_{ind}(\mu, \nu) = \nu^{(1)} \otimes \mu^{(2)}$, where $\nu^{(1)}$ if the first marginal of ν and $\mu^{(2)}$ is the second marginal of μ.*

Given the composition rule $\mathcal{R}_{ind}$, it is then possible to retrieve the heuristic $\mathcal{S}_{\mathcal{R}_{ind}}$, whose transportation cost is lower than the one induced by $\pi = \mu \otimes \nu$, regardless of the power p.

Proposition 2. *Given two probability distributions μ and ν, for any $p \geq 1$ we have that $\mathbb{T}_p(\mathcal{S}_{\mathcal{R}_{ind}}(\mu, \nu)) \leq \mathbb{T}_p(\mu \otimes \nu)$.*

Remark 1. *Albeit the independent composition rule is simplistic, it was shown in [5] that the lift function induced by $\mathcal{R}_{ind}$ returns the optimal transportation plan when μ and ν are both independent.*

3.2.3 The Optimal Composition Rule. Finally, we consider the optimal composition rule, defined via the following identity

$$\mathcal{R}_{opt}(\mu, \nu) \in \underset{\zeta \in \mathcal{I}(\mu,\nu)}{\arg\min} \sum_{i_2} W_p^p(\mu_{|i_2}, \zeta_{|i_2}) \mu^{(2)}_{i_2} + \sum_{j_1} W_p^p(\zeta_{|j_1}, \nu_{|j_1}) \nu^{(1)}_{j_1}. \tag{8}$$

The optimal composition rule is relevant, as it induces an optimal transportation plan. In particular, being able to solve the problem (8) is equivalent to solving the optimal transportation plan between any couple of probability distributions.

Theorem 4. *The function $\mathcal{S}_{\mathcal{R}_{opt}}$ induced by $\mathcal{R}_{opt}$ always returns an optimal transportation plan between the input measures.*

3.3 The Properties of the KR-Heuristics

To conclude the section, we study the theoretical properties of the function $\mathcal{S}_{\mathcal{R}}$. First, we characterise the plan as the minimum of a suitable minimisation problem.

Proposition 3. *Given two probability measures $\mu, \nu \in \mathcal{P}([n]^2)$, for any given composition rule $\mathcal{R}$, let us consider the function $\mathcal{S}_{\mathcal{R}}$ defined as in (6). Then, given $\zeta = \mathcal{R}(\mu, \nu)$, we have that*

$$\mathcal{S}_{\mathcal{R}}(\mu, \nu) \in \underset{\pi \in \Pi_\zeta(\mu,\nu)}{\arg\min} \sum_{i,j} c_{i,j} \pi_{i,j}, \tag{9}$$

where $\Pi_\zeta(\mu, \nu) := \{\pi \in \Pi(\mu, \nu) \text{ s.t. } \sum_{i_1,j_2} \pi_{i_1,i_2,j_1,j_2} = \zeta_{j_1,i_2}\}$.

We now study the convexity properties of the heuristics induced by a composition rule. First, we recall that the convex combination of two composition rules is still a composition rule albeit we have $\lambda \mathcal{S}_{\mathcal{R}} + (1-\lambda)\mathcal{S}_{\mathcal{R}'} \neq \mathcal{S}_{\lambda\mathcal{R}+(1-\lambda)\mathcal{R}'}$, as the following example shows.

Example 1. *Let us consider two probability measures defined as follows. First, μ is supported over $\{(0,0), (0,1), (1,0), (1,1)\}$ and each point in the support is associated with $\frac{1}{4}$ probability. Second, ν is supported over $\{(2,2), (2,3), (3,2), (3,3)\}$ and each point in the support is associated with probability $\frac{1}{4}$. Let us now consider two composition rules, namely $\mathcal{R}$ and $\mathcal{R}'$, such that $\mathcal{R}(\mu, \nu) = \frac{1}{2}(\delta_{(2,0)} + \delta_{(3,1)})$*

and $\mathcal{R}'(\mu,\nu) = \frac{1}{2}(\delta_{(2,1)} + \delta_{(3,0)})$, *hence* $\frac{1}{2}\mathcal{R}(\mu,\nu) + \frac{1}{2}\mathcal{R}'(\mu,\nu) = \frac{1}{4}(\delta_{(2,0)} + \delta_{(3,1)} + \delta_{(2,1)} + \delta_{(3,0)})$. *A simple computation shows that*

$$\mathcal{S}_{\frac{1}{2}\mathcal{R}+\frac{1}{2}\mathcal{R}'}(\mu,\nu) = \frac{1}{4}(\delta_{(0,0,2,2)} + \delta_{(0,1,2,3)} + \delta_{(1,0,3,2)} + \delta_{(1,1,3,3)})$$

and hence $\frac{1}{2}\mathcal{S}_{\mathcal{R}}(\mu,\nu) + \frac{1}{2}\mathcal{S}_{\mathcal{R}'}(\mu,\nu) \neq \mathcal{S}_{\frac{1}{2}\mathcal{R}+\frac{1}{2}\mathcal{R}'}(\mu,\nu)$.

Building on Example 1 and on Proposition 3, we show that the functional $\mathbb{T}_p$ is convex with respect to the composition rules.

Corollary 1. *Given two probability distributions* μ *and* ν, *let* $\mathbb{T}_p(\pi) := \sum_{\boldsymbol{i},\boldsymbol{j}\in[n]^2} ||\boldsymbol{i}-\boldsymbol{j}||_p^p \pi_{\boldsymbol{i},\boldsymbol{j}}$. *For every* $\lambda \in [0,1]$, *we have*

$$\mathbb{T}_p(\mathcal{S}_{\lambda\mathcal{R}+(1-\lambda)\mathcal{R}'}(\mu,\nu)) \leq \mathbb{T}_p\Big(\lambda\mathcal{S}_{\mathcal{R}}(\mu,\nu) + (1-\lambda)\mathcal{S}_{\mathcal{R}'}(\mu,\nu)\Big). \quad (10)$$

Corollary 1 is particularly important when we try to characterise the optimal composition rule. Moreover, going back to Section 3.2.3, we have that the objective and the set of feasible composition rules are convex. To conclude the section, we present two estimates on the Wasserstein distance between the output of two different composition rules, namely $\mathcal{R}$ and $\mathcal{R}'$. In the first result, we quantify how close the two composition rules are in terms of the marginal distance between the outputs, while in the second result, we show that the output of the two composition rules are closer than the output of the two induced heuristic functions.

Theorem 5. *Let* $\mathcal{R}$ *and* $\mathcal{R}'$ *be two composition rules. Given two probability measures* $\mu,\nu \in \mathcal{P}([n]]^2)$, *we have that for any* $p \geq 1$

$$W_p^p(\zeta,\zeta') \leq \sum_{i_2=1}^{n} W_p^p(\zeta_{|i_2},\zeta'_{|i_2})\mu_{i_2}^{(2)} + \sum_{j_1=1}^{n} W_p^p(\zeta_{|j_1},\zeta'_{|j_1})\nu_{j_1}^{(1)},$$

where $\zeta = \mathcal{R}(\mu,\nu)$ *and* $\zeta' = \mathcal{R}'(\mu,\nu)$.

Theorem 5 allows us to bound how close a generic composition rule is from the optimal composition rule. In particular, we have that the difference between a composition rule and the optimal composition rule depends on the conditional distributions of the underlying measures, weighted by their respective marginals. This highlights that deviations at the global level can be traced back to differences in the conditional measures, thus providing a structured way to analyse the stability of composition rules.

Theorem 6. *Given* $p \geq 1$, *let* $\mu,\nu \in \mathcal{P}([n]^2)$ *be two probability measures and let* $\mathcal{R}$ *and* $\mathcal{R}'$ *be two composition rules. Denoted with* $\mathcal{S}_{\mathcal{R}}$ *and* $\mathcal{S}_{\mathcal{R}'}$ *the heuristics induced by* $\mathcal{R}$ *and* $\mathcal{R}'$, *we have that*

$$W_p^p(\mathcal{R}(\mu,\nu),\mathcal{R}'(\mu,\nu)) \leq W_p^p(\mathcal{S}_{\mathcal{R}}(\mu,\nu),\mathcal{S}_{\mathcal{R}'}(\mu,\nu)).$$

Notice that Theorem 6 also acts as a lower bound on how close the two heuristics are. In particular, two different composition rules necessarily induce two different heuristics.

4 A NEURAL NETWORK APPROACH TO OPTIMAL COMPOSITION RULES

In this section, we address the problem of retrieving the optimal composition rule $\mathcal{R}$ via a neural network. Since the space of admissible composition rules is convex and the associated objective functional is convex with respect to $\mathcal{R}$, the optimal composition rule is characterised as the solution of a convex minimisation problem

$$\min_{\zeta\in\mathcal{I}(\mu,\nu)} \sum_{i_2} W_p(\mu_{|i_2},\zeta_{|i_2})\mu_{i_2}^{(2)} + \sum_{j_1} W_p(\zeta_{|j_1},\nu_{|j_1})\nu_{j_1}^{(1)}, \quad (11)$$

for every pair of probability distributions μ and ν. To tackle this problem, we propose a neural network framework designed to approximate the mapping $\mathcal{R}_{\text{opt}} : \mathcal{P}([n]^2)\times\mathcal{P}([n]^2) \longrightarrow \mathcal{P}([n]^2)$, where $\mathcal{P}([n]^2)$ denotes the set of probability measures supported on discrete grids. The optimisation is performed via the method of Augmented Lagrangian Multipliers [22, 42] in order to efficiently enforce the convex constraints. For the sake of simplicity, we consider the case in which we want to retrieve the optimal composition rule associated with the W_1 distance. The case in which we want to approximate the other W_p distances is slightly more delicate, and it is addressed in a dedicated appendix.

Let us then consider the case in which $p = 1$. We therefore want to build a neural network that retrieves the best composition rule associated with the Wasserstein distance W_1. Owing to the properties of the W_1 distance, we rewrite the objective in (11) as

$$\mathcal{L}_1(\zeta) = \sum_{i_2,t=1}^{n}\Big|\sum_{l=1}^{t}\mu_{l,i_2} - \sum_{l=1}^{t}\zeta_{l,i_2}\Big|\mu_{j_2}^{(2)} + \sum_{j_1,t=1}^{n}\Big|\sum_{l=1}^{t}\nu_{j_1,l} - \sum_{l=1}^{t}\zeta_{j_1,l}\Big|\nu_{j_1}^{(1)}.$$

First, notice that $\mathcal{L}$ is convex with respect to ζ. We now need to enforce the constraint $\zeta \in \mathcal{I}(\mu,\nu)$ in the objective function. To do that, we follow the approach outlined in [22] and resort to a Lagrangian formulation and add the constraint to the loss function. As a result, we obtain the following loss function

$$\begin{aligned}\tilde{\mathcal{L}}_1(\zeta) = &\sum_{j=1}^{n}\left(\sum_{T=1}^{n}\Big|\sum_{i=1}^{T}\mu_{i,j} - \sum_{k=1}^{T}\zeta_{k,j}\Big| + \lambda_j^{(1)}\left(\sum_{k=1}^{n}\zeta_{k,j} - \sum_{i=1}^{n}\mu_{i,j}\right)\right)\\ &+ \sum_{k=1}^{n}\left(\sum_{T=1}^{n}\Big|\sum_{l=1}^{T}\nu_{k,l} - \sum_{j=1}^{T}\zeta_{k,j}\Big| + \lambda_k^{(2)}\left(\sum_{j=1}^{n}\zeta_{k,j} - \sum_{l=1}^{n}\nu_{k,l}\right)\right)\\ &+ \rho\left(\sum_{j=1}^{n}\Big|\sum_{k=1}^{n}\zeta_{k,j} - \sum_{i=1}^{n}\mu_{i,j}\Big|^2 + \sum_{k=1}^{n}\Big|\sum_{j=1}^{n}\zeta_{k,j} - \sum_{l=1}^{n}\nu_{k,l}\Big|^2\right)\end{aligned}$$

where (i) $\lambda^{(1)}$ and $\lambda^{(2)}$ are the Lagrangian multiplier associated with the constraints over ζ and (ii) $\rho > 0$ is a penalty parameter, whose value can be fixed arbitrarily. Notice that $\tilde{\mathcal{L}}$ is still convex with respect to ζ. Moreover, the minimiser of $\tilde{\mathcal{L}}$ is a minimiser of $\mathcal{L}$, hence we can adopt $\tilde{\mathcal{L}}$ as a loss function to define a neural network able to infer the optimal composition rule.

5 NUMERICAL EXPERIMENTS

In this section, we assess the performance of the Neural Network that builds the optimal transportation plan through the optimal composition rule generated by our generalised Knothe-Rosenblatt rearrangements. We name our Neural Network KRo-Net. Comparisons are made with representative baselines on multiple benchmark datasets, focusing on accuracy and computational efficiency.

5.1 Experimental settings

Benchmark datasets. We conduct experiments on multiple public datasets from the domains of computer vision and medical imaging to assess our method. All images are preprocessed by normalising the total pixel values to 1, thus representing discrete probability distributions. For each dataset, we split images belonging to different classes evenly between sources and targets. The benchmark datasets are as follows: (i) MNIST [20]: The MNIST dataset consists of grayscale images of handwritten digits (0–9), each of size

28×28. (ii) USPS [29]: A handwritten digit dataset consisting of 7,291 training images and 2,007 test images. (iii) BraTS2020 [27]: The BraTS2020 dataset contains multi-modal brain tumor magnetic resonance imaging (MRI) scans from 369 training and 125 validation subjects. (iv) Brain MRI Dataset [37]: 7,017 brain MRI images in four categories (glioma, meningioma, pituitary, no tumor).

Baselines. To validate the effectiveness of our KRo-Net, we compare it against three representative baselines: (i) Knothe–Rosenblatt rearrangement (KR) [32, 43]: A heuristic staged matching method that aligns row and column cumulative distributions to generate transport plans. (ii) Sinkhorn algorithm [17]: An entropy-regularised optimal transport algorithm that iteratively computes approximate solutions, striking a balance between accuracy and efficiency. (iii) Sinkhorn-Net [54]: A neural network-based method that learns potential functions for distributions and derives transport matrices via end-to-end optimisation. These baselines collectively cover heuristic, optimisation-based, and neural network-based strategies for solving optimal transport.

Evaluation metrics. We adopt the following metrics to ensure a comprehensive evaluation of different methods: (i) Wasserstein distance (defined as in Definition 2): The two-dimensional Wasserstein distance between the transported distribution and the target distribution. Smaller distances imply closer alignment with the target distribution. (ii) Computational complexity: The runtime and memory consumption of each method are recorded to assess computational efficiency and scalability.

Implementation details. Our proposed method is implemented as a multi-layer perceptron with two hidden layers of 128 and 64 units and a Sigmoid output producing the intermediate measure. We intentionally adopt a basic network architecture to show that our approach achieves strong performance without resorting to sophisticated networks. For baselines, we use the hyper-parameters reported in the original papers. When unavailable, we perform simple tuning. All methods are re-implemented in the same PyTorch 2.9.0 framework for consistency. Results are averaged over 100 runs. Experiments are conducted on an Nvidia GeForce RTX 5090 GPU. The code is available at https://github.com/1291413/KRo-Net.git.

5.2 Performance Comparison

Owing to the different nature of the baselines considered, we study the accuracy and the computational complexity separately.

Accuracy. To evaluate the effectiveness of our proposed method in learning accurate transport plans, we first assess the accuracy of the methods. We measure the accuracy as the Wasserstein distance achieved by four methods on MNIST, USPS, BraTS2020, and Brain MRI datasets (see Table 2). Several key observations can be drawn: (i) across all datasets, our method consistently achieves the smallest Wasserstein distance, often several orders of magnitude lower than baselines. This indicates that our approach is able to approximate the optimal transport solution with much higher precision; (ii) classical solvers such as KR and Sinkhorn suffer from large deviations, especially on high-dimensional medical imaging datasets (e.g., BraTS2020 and Brain MRI), where their Wasserstein values remain at least three magnitudes higher than ours; and (iii) Sinkhorn-Net, though designed for learning-based approximation, exhibits unstable behaviour and fails to converge to sufficiently small transport costs, highlighting the advantage of our principled formulation.

Table 1: Comparisons of four OT methods on the benchmark datasets. Results are reported as mean ± standard deviation over 100 runs, using Accuracy as evaluation metric. Best results are in bold, second best are underlined.

Dataset	Method	Wasserstein Distance
MNIST	KRo-Net(ours)	$\underline{2.07 \times 10^{-8}} \pm \underline{8.40 \times 10^{-9}}$
	KR	$1.205 \pm 3.35 \times 10^{-1}$
	Sinkhorn	$3.31 \times 10^{-1} \pm 9.30 \times 10^{-2}$
	Sinkhorn-Net	2.837 ± 1.442
USPS	KRo-Net(ours)	$\underline{1.0 \times 10^{-10}} \pm \underline{3.0 \times 10^{-10}}$
	KR	$7.72 \times 10^{-1} \pm 7.2 \times 10^{-2}$
	Sinkhorn	$3.69 \times 10^{-1} \pm 9.3 \times 10^{-2}$
	Sinkhorn-Net	$1.401 \pm 6.88 \times 10^{-1}$
BraTS2020	KRo-Net(ours)	$\underline{2.12 \times 10^{-8}} \pm \underline{3.0 \times 10^{-9}}$
	KR	$1.029 \pm 8.86 \times 10^{-1}$
	Sinkhorn	9.13 ± 4.932
	Sinkhorn-Net	9.170 ± 8.826
Brain MRI	KRo-Net(ours)	$\underline{5.99 \times 10^{-8}} \pm \underline{2.46 \times 10^{-8}}$
	KR	$9.76 \times 10^{-1} \pm 1.15 \times 10^{-1}$
	Sinkhorn	6.948 ± 2.014
	Sinkhorn-Net	$1.2296 \times 10^{1} \pm 3.002$

Furthermore, to provide a more intuitive understanding, Figure 3 visualises the source-to-target distribution alignment on Brain MRI slices. While baselines produce blurry or distorted mappings between source and target distributions, our method yields sharp and semantically consistent transport, preserving tumor boundaries and brain structures. This qualitative comparison further validates that minimising Wasserstein distance in our framework not only improves theoretical transport accuracy but also translates into superior visual alignment in real medical data.

Complexity and Memory Usage. We now evaluate the computational complexity of the two Neural Networks baselines (see Table 2). Indeed, comparing the efficiency our approach and Sinkhorn-Net with the classic Sinkhorn and the KR rearrangement is unfair, as the latter two are closed form methods and thus do not entail a Neural Network training. To assess the computational complexity we consider two main indicators: the runtime and the memory usage. On both metrics, our approach outclasses the Sinkhorn net on basically all the considered datasets. In particular, our method consistently maintains low runtime on large-scale data such as BraTS2020 and Brain MRI, where learning-based SinkhornNet suffers from extremely high computational overhead. For example, on Brain MRI, SinkhornNet requires more than 30,000 seconds on average, while our method takes a few hundred seconds. This demonstrates that our framework scales far more effectively in practice. Likewise, we observe that our approach uses moderate memory compared to traditional solvers. On MNIST and USPS, memory consumption remains within a few hundred MB, significantly lower than SinkhornNet, which can exceed several GB due to deep network parameterisation. Similar considerations can be drawn for larger datasets such as BraTS2020 and Brain MRI.

Table 2: Comparisons of the Neural Network based methods on the selected datasets. Results are reported as mean ± standard deviation over 100 runs, using accuracy, runtime, and memory as evaluation metrics. Best results are in bold.

Dataset	Method	Wasserstein Distance	Time (s)	Memory (MB)
MNIST	KRo-Net(ours)	$\underline{2.07 \times 10^{-8} \pm 8.40 \times 10^{-9}}$	$\underline{1.644\,43 \times 10^{2} \pm 1.056\,03 \times 10^{2}}$	$1.355\,740 \times 10^{3} \pm \underline{3.963}$
	Sinkhorn-Net	2.837 ± 1.442	$4.208\,766 \times 10^{3} \pm 1.128\,492 \times 10^{3}$	$\underline{1.559\,46 \times 10^{2}} \pm 5.4386 \times 10^{1}$
USPS	KRo-Net(ours)	$\underline{1.0 \times 10^{-10} \pm 3.0 \times 10^{-10}}$	$\underline{7.442\,96 \times 10^{2} \pm 3.007\,45 \times 10^{2}}$	$8.859\,77 \times 10^{2} \pm 4.884\,41 \times 10^{2}$
	Sinkhorn-Net	$1.401 \pm 6.88 \times 10^{-1}$	$4.152\,837 \times 10^{3} \pm 2.365\,790 \times 10^{3}$	$\underline{1.065\,55 \times 10^{2}} \pm \underline{3.3642 \times 10^{1}}$
BraTS2020	KRo-Net(ours)	$\underline{2.12 \times 10^{-8} \pm 3.0 \times 10^{-9}}$	$\underline{2.473\,65 \times 10^{2} \pm 6.759}$	$1.611\,887\,7 \times 10^{4} \pm \underline{1.602\,596 \times 10^{3}}$
	Sinkhorn-Net	9.170 ± 8.826	$65\,792.781 \pm 1656.426$	$\underline{4236.043} \pm 4055.556$
Brain MRI	KRo-Net(ours)	$\underline{5.99 \times 10^{-8} \pm 2.46 \times 10^{-8}}$	$\underline{3.923\,79 \times 10^{2} \pm 1.186\,64 \times 10^{2}}$	$4.710\,57 \times 10^{3} \pm \underline{2.523\,86 \times 10^{2}}$
	Sinkhorn-Net	$1.2296 \times 10^{1} \pm 3.002$	$3.019\,344\,7 \times 10^{4} \pm 7.562\,55 \times 10^{2}$	$\underline{3.572\,436 \times 10^{3}} \pm 2.404\,801 \times 10^{3}$

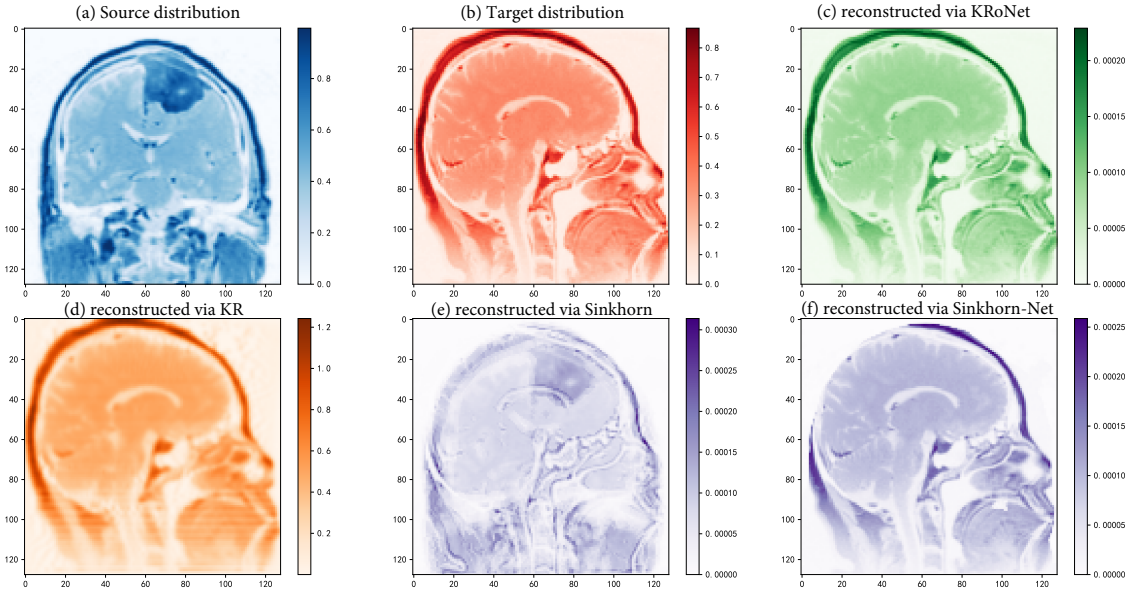


Figure 3: Source-to-target alignment on Brain MRI slices: (a) source distribution, (b) target distribution, (c) reconstruction by KRo-Net (our approach), (d) the Knothe-Rosenblatt heuristic, (e) Sinkhorn, and (f) Sinkhorn-Net.

Conclusions. Taken together, these results highlight that our method not only achieves the most accurate Wasserstein approximation (orders of magnitude better), but also offers a reduced computational complexity, making it appealing to real-world applications, such as medical imaging. In particular, we notice that even using a simple Neural Network, our approach outclasses more sophisticated and recognised neural structures such as Sinkhorn-Net.

6 CONCLUSION AND FUTURE WORKS

In this work, we introduced a novel class of heuristic transportation plans that generalises the Knothe–Rosenblatt rearrangement [32, 43]. At the core of our framework lies the notion of a *composition rule*, a mapping that, given two distributions μ and ν, produces an auxiliary distribution ζ that induces a transportation plan. This perspective recovers the Knothe–Rosenblatt rearrangement and the optimal transport plan, while also enabling the construction of new heuristic couplings. We established theoretical guarantees by showing that every composition rule corresponds to the solution of a suitable class of optimal transport problems, together with explicit error bounds. Finally, we demonstrated how our framework can be adapted into an efficient neural network model that approximates optimal transportation plans, serving as an oracle for otherwise costly computations and offering promising applications in tasks such as clustering, classification, and medical image analysis. In our future works, we will apply this formalism to higher-dimensional problems and to other Optimal Transport problems.

ACKNOWLEDGMENTS

ZG acknowledges support from the High-level Talent Project of Xiamen University of Technology (YKJ24016R). ZC acknowledges support from the Natural Science Foundation of Xiamen, China (Grant No. 3502Z202571066).

REFERENCES

[1] Priyank Agrawal, Eric Balkanski, Vasilis Gkatzelis, Tingting Ou, and Xizhi Tan. 2022. Learning-augmented mechanism design: Leveraging predictions for facility location. In *Proceedings of the 23rd ACM Conference on Economics and Computation*. 497–528.

[2] Sina Akbari, Luca Ganassali, and Negar Kiyavash. 2023. Causal discovery via monotone triangular transport maps. In *NeurIPS 2023 Workshop Optimal Transport and Machine Learning*. NeurIPS, New Orleans.

[3] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. 2017. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances in neural information processing systems* 30 (2017).

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. *Proceedings of Machine Learning Research* 70 (06–11 Aug 2017), 214–223.

[5] Gennaro Auricchio. 2024. On the pythagorean structure of the optimal transport for separable cost functions. *Rendiconti Lincei* 34, 4 (2024), 745–771.

[6] Gennaro Auricchio, Federico Bassetti, Stefano Gualandi, and Marco Veneroni. 2018. Computing Kantorovich-Wasserstein Distances on d-dimensional histograms using $(d+1)$-partite graphs. In *Advances in Neural Information Processing Systems*. 5793–5803.

[7] Gennaro Auricchio, Gabriele Loli, and Marco Veneroni. 2025. On the computation of the infinity Wasserstein distance and the Wasserstein Projection Problem. *J. Comput. Appl. Math.* (2025), 117025.

[8] Gennaro Auricchio and Jie Zhang. 2025. Leveraging Optimal Transport to Design Optimal Mechanisms for the Facility Location Problem. *ACM Trans. Econ. Comput.* (2025). https://doi.org/10.1145/3757106

[9] Gennaro Auricchio, Jie Zhang, and Mengxiao Zhang. 2024. Extended ranking mechanisms for the m-capacitated facility location problem in bayesian mechanism design. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 87–95.

[10] Julio Backhoff, Mathias Beiglbock, Yiqing Lin, and Anastasiia Zalashko. 2017. Causal transport in discrete time and applications. *SIAM Journal on Optimization* 27, 4 (2017), 2528–2562.

[11] Ricardo Baptista, Youssef Marzouk, and Olivier Zahm. 2024. On the representation and learning of monotone triangular transport maps. *Foundations of Computational Mathematics* 24, 6 (2024), 2063–2108.

[12] Zohar Barak, Anupam Gupta, and Inbal Talgam-Cohen. 2024. MAC advice for facility location mechanism design. *Advances in Neural Information Processing Systems* 37 (2024), 129564–129604.

[13] Federico Bassetti, Stefano Gualandi, and Marco Veneroni. 2020. On the Computation of Kantorovich-Wasserstein Distances between 2D-Histograms by Uncapacitated Minimum Cost Flows. *SIAM, Journal on Optimization* 30, 3 (2020), 2441–2469.

[14] Nicolas Bonneel and David Coeurjolly. 2019. Spot: sliced partial optimal transport. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–13.

[15] Yann Brenier. 1991. On the translocation of masses. *Communications on pure and applied mathematics* 44, 4 (1991), 375–417.

[16] José A. Carrillo, Marco Di Francesco, and Corrado Lattanzio. 2007. Contractivity and asymptotics in Wasserstein metrics for viscous nonlinear scalar conservation laws. *Journal of Differential Equations* 10, 2 (June 2007), 277–292.

[17] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*. 2292–2300.

[18] Marco Cuturi and Arnaud Doucet. 2014. Fast Computation of Wasserstein Barycenters. *Proceedings of Machine Learning Research* 32, 2 (22–24 Jun 2014), 685–693.

[19] Constantinos Daskalakis, Alan Deckelbaum, and Christos Tzamos. 2013. Mechanism design via optimal transport. In *Proceedings of the fourteenth ACM conference on Electronic commerce*. 269–286.

[20] Li Deng. 2012. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine* 29, 6 (2012), 141–142. https://doi.org/10.1109/MSP.2012.2211477

[21] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2017. Density estimation using Real NVP. In *International Conference on Learning Representations*.

[22] Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David Parkes, and Sai Srivatsa Ravindranath. 2019. Optimal auctions through deep learning. In *International Conference on Machine Learning*. PMLR, 1706–1715.

[23] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A. Poggio. 2015. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*. 2053–2061.

[24] Andrew V. Goldberg, Éva Tardos, and Robert Tarjan. 1989. *Network flow algorithm*. Technical Report. Cornell University Operations Research and Industrial Engineering.

[25] Yannai A Gonczarowski. 2018. Bounding the menu-size of approximately optimal auctions via optimal-transport duality. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. 123–131.

[26] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems* 30 (2017).

[27] Theophraste Henry, Alexandre Carré, Marvin Lerousseau, Théo Estienne, Charlotte Robert, Nikos Paragios, and Eric Deutsch. 2020. Brain tumor segmentation with self-ensembled, deeply-supervised 3D U-net neural networks: a BraTS 2020 challenge solution. In *International MICCAI Brainlesion Workshop*. Springer, 327–339.

[28] Pierre Henry-Labordere and Tiphaine Monedero. 2022. From Knothe-Rosenblatt Rearrangement to Distribution Mapping for Gaussian Fixed-Income Models. *Available at SSRN 4232008* (2022).

[29] J. J. Hull. 1994. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16, 5 (1994), 550–554. https://doi.org/10.1109/34.291440

[30] Priyank Jaini, Kira A Selby, and Yaoliang Yu. 2019. Sum-of-squares polynomial flow. In *International Conference on Machine Learning*. PMLR, 3009–3018.

[31] Leonid V. Kantorovich. 1960. Mathematical methods of organizing and planning production. *Management science* 6, 4 (1960), 366–422.

[32] Herbert Knothe. 1957. Contributions to the theory of convex bodies. *Michigan Math. J.* 4, 1 (1957), 39–52. https://doi.org/10.1307/mmj/1028990175

[33] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. 2020. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence* 43, 11 (2020), 3964–3979.

[34] Haibin Ling and Kazunori Okada. 2007. An efficient Earth Mover's Distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 5 (2007), 840–853.

[35] John N. Mather. 1991. Action minimizing invariant measures for positive definite Lagrangian systems. *Mathematische Zeitschrift* 207, 2 (1991), 169–208. http://eudml.org/doc/174264

[36] Gaspard Monge. 1781. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris* (1781).

[37] Msoud Nickparvar. 2021. Brain Tumor MRI Dataset. https://doi.org/10.34740/KAGGLE/DSV/2645886

[38] James B Orlin. 1993. A faster strongly polynomial minimum cost flow algorithm. *Operations research* 41, 2 (1993), 338–350.

[39] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. 2021. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research* 22, 57 (2021), 1–64.

[40] Ofir Pele and Michael Werman. 2009. Fast and robust Earth Mover's Distances. In *Computer vision, 2009 IEEE 12th international conference on*. IEEE, 460–467.

[41] Henning Petzka, Asja Fischer, and Denis Lukovnikov. 2018. On the regularization of Wasserstein GANs. In *International Conference on Learning Representations*.

[42] R Tyrrell Rockafellar. 1974. Augmented Lagrange multiplier functions and duality in nonconvex programming. *SIAM Journal on Control* 12, 2 (1974), 268–285.

[43] Murray Rosenblatt. 1952. Remarks on a multivariate transformation. *The annals of mathematical statistics* 23, 3 (1952), 470–472.

[44] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 1998. Metric for distributions with applications to image databases. *Proceedings of the IEEE International Conference on Computer Vision* (02 1998), 59–66. https://doi.org/10.1109/ICCV.1998.710701

[45] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 1998. A metric for distributions with applications to image databases. In *Computer Vision, 1998. Sixth International Conference on*. IEEE, 59–66.

[46] Filippo Santambrogio. 2015. Optimal transport for applied mathematicians. *Birkäuser, NY* (2015).

[47] Alessandro Scagliotti and Sara Farinelli. 2025. Normalizing flows as approximations of optimal transport maps via linear-control neural ODEs. *Nonlinear Analysis* 257 (2025), 113811.

[48] Kulin Shah, Amit Deshpande, and Navin Goyal. 2022. Learning and Generalization in Overparameterized Normalizing Flows. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 9430–9504.

[49] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. 2015. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 66.

[50] Keju Tang, Xiaoliang Wan, and Qifeng Liao. 2020. Deep density estimation via invertible block-triangular mapping. *Theoretical and Applied Mechanics Letters* 10, 3 (2020), 143–148.

[51] Cédric Villani. 2008. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.

[52] Jiqing Wu, Zhiwu Huang, Janine Thoma, Dinesh Acharya, and Luc Van Gool. 2018. Wasserstein divergence for gans. In *Proceedings of the European conference on computer vision (ECCV)*. 653–668.

[53] Chenyang Xu and Pinyan Lu. 2022. Mechanism design with predictions. (2022).

[54] Jie Xu, Chaozhuo Li, Feiran Huang, Zhoujun Li, Xing Xie, and Philip S. Yu. 2024. Sinkhorn Distance Minimization for Adaptive Semi-Supervised Social Network Alignment. *IEEE Transactions on Neural Networks and Learning Systems* 35, 10 (2024), 13340–13353. https://doi.org/10.1109/TNNLS.2023.3267126