Hierarchical Hybrid Sliced Wasserstein: A Scalable Metric for Heterogeneous Joint Distributions

Khai Nguyen Nhat Ho

The University of Texas at Austin
April 24, 2024

Abstract

Sliced Wasserstein (SW) and Generalized Sliced Wasserstein (GSW) have been widely used in applications due to their computational and statistical scalability. However, the SW and the GSW are only defined between distributions supported on a homogeneous domain. This limitation prevents their usage in applications with heterogeneous joint distributions with marginal distributions supported on multiple different domains. Using SW and GSW directly on the joint domains cannot make a meaningful comparison since their homogeneous slicing operator i.e., Radon Transform (RT) and Generalized Radon Transform (GRT) are not expressive enough to capture the structure of the joint supports set. To address the issue, we propose two new slicing operators i.e., Partial Generalized Radon Transform (PGRT) and Hierarchical Hybrid Radon Transform (HHRT). In greater detail, PGRT is the generalization of Partial Radon Transform (PRT), which transforms a subset of function arguments non-linearly while HHRT is the composition of PRT and multiple domain-specific PGRT on marginal domain arguments. By using HHRT, we extend the SW into Hierarchical Hybrid Sliced Wasserstein (H2SW) distance which is designed specifically for comparing heterogeneous joint distributions. We then discuss the topological, statistical, and computational properties of H2SW. Finally, we demonstrate the favorable performance of H2SW in 3D mesh deformation, deep 3D mesh autoencoders, and datasets comparison.

1 Introduction

Optimal transport [56, 48] is a powerful mathematical tool for machine learning, statistics, and data sciences. As an example, Wasserstein distance [48], defined as the optimal transportation cost between two distributions, has been used successfully in many areas of machine learning and statistics, such as generative modeling on images [2, 54], representation learning [36], vocabulary learning [59], and so on. Despite being accepted as an effective distance, Wasserstein distance has been widely known as a computationally expensive distance. In particular, when comparing two distributions that have at most n supports, the time complexity and the memory complexity of the Wasserstein distance scale with the order of $\mathcal{O}(n^3 \log n)$ [46] and $\mathcal{O}(n^2)$ respectively. In addition, the Wasserstein distance requires more samples to approximate a continuous distribution with its empirical distribution in high dimension since its sample complexity is of the order of $\mathcal{O}(n^{-1/d})$ [20], where n is the sample size and d is the number of dimensions. Therefore, Wasserstein distance is not statistically and computationally scalable, especially in high dimensions.

Along with entropic regularization [17] which can reduce the time complexity and memory complexity of computing optimal transport to $\mathcal{O}(n^2)$ and $\mathcal{O}(n^2)$ in turn, sliced Wasserstein (SW) distance [11] is one alternative approach for the original Wasserstein distance. The key benefit of the SW distance is that it scales the order $\mathcal{O}(n \log n)$ and $\mathcal{O}(n)$ in terms of time and memory respectively. The reason

behind that fast computation is the closed-form solution of optimal transport in one dimension. To leverage that closed-form, sliced Wasserstein utilizes Radon Transform [24] (RT) to transform a high-dimensional distribution to its one-dimensional projected distributions, then the final distance is calculated as the average of all one-dimensional Wasserstein distance. By doing that, the SW distance has a very fast sample complexity i.e., $\mathcal{O}(n^{-1/2})$, which makes it computationally and statistically scalable in any dimension. Therefore, the SW distance has been applied successfully in various domains of applications including generative models [19], domain adaptation [33], clustering [28], 3D shapes [31, 30], gradient flows [35, 8], Bayesian inference computation [38, 60], texture synthesis [23], and many other tasks.

Despite being useful, the SW distance is not as flexible as the Wasserstein distance in terms of choosing the ground metric. In greater detail, the number of ground metrics in one dimension is limited, especially ground metrics that lead to the closed-form solution. As a result, the role of capturing the structure of interested distributions belongs to the slicing/projecting operators. To generalize RT to non-linear projection, generalized Radon Transform (GRT) is introduced in [3] with circular projection [29], polynomial projection [52], and so on. With GRT, Generalized Sliced Wasserstein (GSW) distance is proposed in [27]. In addition, a line of works on developing sliced Wasserstein variants on different manifolds such as hyper-sphere [6, 55, 50, 51], hyperbolic manifolds [7], the manifold of symmetric positive definite matrices [10], general manifolds and graphs [53]. In those works, special variants of GRT are proposed.

Although the SW has become more effective on multiple domains, no SW variant is designed specifically for heterogeneous joint distributions i.e., joint distributions that have marginals supported on different domains, except for the product of Hadamard manifolds [9]. It is worth noting that marginal domains of heterogeneous joint distributions can be any metric space and are not necessary manifolds. Heterogeneous joint distributions appear in many applications, e.g., domain adaptation domains [15, 4], comparing datasets with labels [1], 3D shape deformation [30], and so on. In this case, Wasserstein distance can be adapted by using a mixed ground metric, i.e., a weighted sum of metrics on domains [15, 1]. In contrast to the Wasserstein distance, the adaptation of SW has not been well-investigated. Using GSW directly with one type defining function for all marginals cannot separate the information within and among groups of arguments.

Contribution: In this work, we tackle the challenge of designing a sliced Wasserstein variant for heterogeneous joint distributions. In summary, our main contributions are three-fold:

1. We first extend the partial Radon Transform to the partial generalized Radon Transform (PGRT) to inject non-linearity into local transformation. We discuss the injectivity of PGRT for some choices of defining functions. We then propose a novel slicing operator for heterogeneous joint distributions, named Hierarchical Hybrid Radon Transform (HHRT). In particular, HHRT is a hierarchical transformation that first applies partial generalized Radon Transform with different defining functions on arguments of each marginal to gather marginal information, then applies partial Radon Transform on the joint transformed arguments to gather information among marginals. We show that HHRT is injective as long as the partial generalized Radon Transform is injective.

2. We propose Hierarchical Hybrid Sliced Wasserstein (H2SW) which is a novel metric for comparing heterogeneous joint distributions by utilizing the HHRT. Moreover, we investigate the topological properties, statistical properties, and computational properties of H2SW. In particular, we show that H2SW is a valid metric on the space of distribution over the joint space, H2SW does not suffer

from the curse of dimensionality and enjoys the same computational scalability as SW distance.

3. A 3D mesh can be effectively represented by a point-cloud and corresponding surface normal vectors. Therefore, it can be seen as an empirical heterogeneous joint distribution. We conduct experiments on optimization-based 3D mesh deformation and deep 3D mesh autoencoder to show the favorable performance of H2SW compared to SW and GSW. Moreover, we also illustrate that H2SW can also provide a meaningful comparison for probability distributions on the product of Hadamand manifolds by conducting experiments on dataset comparison.

Organization. We first provide some preliminaries on SW distance, GSW distance, and joint Wasserstein distance in Section 2. We then define the hierarchical hybrid Radon transform and hierarchical hybrid sliced Wasserstein distance s in Section 3. Section 4 contains experiments on 3D mesh deformation, deep 3D mesh autoencoder, and datasets comparison. We conclude the paper in Section 5. Finally, we defer the proofs of key results, and additional materials in the Appendices.

2 Preliminaries

Wasserstein distance. For $p \geq 1$, the Wasserstein-p distance [57, 48] between two distributions $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, where $\mathcal{X}$ and $\mathcal{Y}$ are subset of $\mathbb{R}^d$ and they share a ground metric $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, is defined as:

$$\mathrm{W}_p^p(\mu, \nu; c) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)^p d\pi(x, y), \tag{1}$$

where $\Pi(\mu, \nu) := \Big\{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})\} | \int_{\mathcal{Y}} d\pi(x, y) = \mu(x), \int_{\mathcal{X}} d\pi(x, y) = \nu(y)\Big\}$. When μ and ν are discrete with at most n supports, the time complexity and the space complexity of the Wasserstein distance is $\mathcal{O}(n^3 \log n)$ and $\mathcal{O}(n^2)$ in turn which are very expensive. Therefore, sliced Wasserstein is proposed as an alternative solution. We first review the

Radon Transform [24] The *Radon Transform* $\mathcal{R} : \mathbb{L}_1(\mathbb{R}^d) \to \mathbb{L}_1\left(\mathbb{R} \times \mathbb{S}^{d-1}\right)$ is defined as:

$$(\mathcal{R}f)(t, \theta) = \int_{\mathbb{R}^d} f(x)\delta(t - \langle x, \theta\rangle)dx. \tag{2}$$

Radon Transform defines a linear bijection [24]. Given a projecting direction θ, $(\mathcal{R}f)(\cdot, \theta)$ is an one-dimensional function. With Radon Transform, we can now define the sliced Wasserstein distance.

Sliced Wasserstein distance. For $p \geq 1$, the *Sliced Wasserstein (SW)* distance [11] of p-th order between two distributions $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ with an one-dimensional ground metric $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is defined as follow:

$$\mathrm{SW}_p^p(\mu, \nu; c) = \mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^{d-1})}[\mathrm{W}_p^p(\mathcal{R}_\theta\sharp\mu, \mathcal{R}_\theta\sharp\nu; c)], \tag{3}$$

where $\mathcal{R}_\theta\sharp\mu$ and $\mathcal{R}_\theta\sharp\nu$ are the one-dimensional push-forward distributions created by applying Radon Transform (RT) [24] on the pdf of μ and ν with the projecting direction θ. The computational benefit of SW distance comes from the closed-form solution when the one-dimensional ground metric $c(x, y) = h(x - y)$ for h is a strictly convex function: $\mathrm{W}_p^p(\mathcal{R}_\theta\sharp\mu, \mathcal{R}_\theta\sharp\nu; c) = \int_0^1 c\left(F^{-1}_{\mathcal{R}_\theta\sharp\mu}(z), F^{-1}_{\mathcal{R}_\theta\sharp\nu}(z)\right)^p dz$, where $F^{-1}_{\mathcal{R}_\theta\sharp\mu}$ and $F^{-1}_{\mathcal{R}_\theta\sharp\nu}$ are inverse CDF of $\mathcal{R}_\theta\sharp\mu$ and $\mathcal{R}_\theta\sharp\nu$ respectively. When μ and ν are discrete

with at most n supports, the time complexity and the space complexity of the closed-form is $\mathcal{O}(n \log n)$ and $\mathcal{O}(n)$ respectively.

Generalized Radon Transform and Sliced Wasserstein distance. To generalize RT to non-linear operator, the *Generalized Radon Transform (GRT)* was proposed [3]. Given a defining function [27] $g : \mathbb{R}^d \times \Omega \to \mathbb{R}$, the Generalized Radon Transform [3] $\mathcal{GR} : \mathbb{L}_1(\mathbb{R}^d) \to \mathbb{L}_1(\mathbb{R} \times \Omega)$ is defined as: $(\mathcal{GR}f)(t, \theta) = \int_{\mathbb{R}^d} f(x)\delta(t - g(x, \theta))dx$. For example, we can have the circular function [29], i.e., $g(x, \theta) = \|x - r\theta\|_2$ for $r \in \mathbb{R}^+$ and $\theta \in \Omega := \mathbb{S}^{d-1}$, homogeneous polynomials with an odd degree [52] (m), i.e., $g(x, \theta) = \sum_{|\alpha|=m} \theta_\alpha x^\alpha$ with $\alpha = (\alpha_1, \ldots, \alpha_{d_\alpha}) \in \mathbb{N}^{d_\alpha}$, $|\alpha| = \sum_{i=1}^{d_\alpha} \alpha_i$, $x^\alpha = \prod_{i=1}^{d_\alpha} x_i^{\alpha_i}$, $\Omega = \mathbb{S}^{d_\alpha}$, and so on. Using GRT, the *Generalized Sliced Wasserstein (GSW)* distance is introduced in [27], which is formally defined as :

$$\text{GSW}_p^p(\mu, \nu; c, g) = \mathbb{E}_{\theta \sim \mathcal{U}(\mathbb{S}^{d-1})}[\text{W}_p^p(\mathcal{GR}_\theta^{g\sharp}\mu, \mathcal{GR}_\theta^{g\sharp}\nu; c)]. \tag{4}$$

It is worth noting that the injectivity of GRT is required to have the identity of indiscernible GSW.

Heterogeneous joint distributions comparison. We are given two joint distributions $\mu(x_1, x_2) \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ and $\nu(y_1, y_2) \in \mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_2)$ where X_1 are Y_1 share a ground metric $c_1 : \mathcal{X}_1 \times \mathcal{Y}_1 \to \mathbb{R}^+$ and X_2 are Y_2 share a ground metric $c_2 : \mathcal{X}_2 \times \mathcal{Y}_2 \to \mathbb{R}^+$ with $(c_1 \neq c_2)$. In this case, previous works utilize the joint distribution Wasserstein distance [15, 1] to compare μ and ν:

$$\text{W}_p^p(\mu, \nu; c_1, c_2) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}_1 \times \mathcal{Y}_2} (c_1(x_1, y_1)^p + c_2(x_2, y_2)^p) d\pi(x_1, x_2, y_1, y_2), \tag{5}$$

where $\Pi(\mu, \nu) := \Big\{\pi \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}_1 \times \mathcal{Y}_2)\} | \int_{\mathcal{Y}_1 \times \mathcal{Y}_2} d\pi(x_1, x_2, y_1, y_2) = \mu(x_1, x_2)$, $\int_{\mathcal{X}_1 \times \mathcal{X}_2} d\pi(x_1, x_2, y_1, y_2) = \nu(y_1, y_2)\Big\}$. We can easily extend the definition to joint distributions with more than two marginals (see Appendix B). In contrast to the Wasserstein distance, there is no variant of SW that is designed specifically for this case. SW variants can still be used by treating $\mathcal{X}_1 \times \mathcal{X}_2$ and $\mathcal{Y}_1 \times \mathcal{Y}_2$ as homogeneous spaces $\mathcal{X}$ and $\mathcal{Y}$ which share the same Radon Transform variant and one-dimensional ground metric c. However, that approach cannot differentiate the difference between $\mathcal{X}_1$ and $\mathcal{X}_2$, and leverage the hierarchical structure, i.e., inside and among marginals.

3 Hierarchical Hybrid Sliced Wasserstein Distance

In this section, we propose the Hierarchical Hybrid Radon Transform (HHRT) which first applies P(G)RT on each marginal argument to gather each marginal information, then applies PRT on the joint transformed arguments from all marginals to gather information among marginals. After that, we introduce Hierarchical Hybrid Sliced Wasserstein distance by using HHRT as the slicing operator.

3.1 Hierarchical Hybrid Radon Transform

We first introduce the first building block in HHRT, i.e., Partial Generalized Radon Transform (PGRT).

Definition 1 (Partial Generalized Radon Transform)**.** *Given a defining function $g : \mathbb{R}^{d_1} \times \Omega \to \mathbb{R}$, Partial Generalized Radon Transform $\mathcal{PGR} : \mathbb{L}_1(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}) \to \mathbb{L}_1\left(\mathbb{R} \times \Omega \times \mathbb{R}^{d_2}\right)$ is defined as:*

$$(\mathcal{PGR}f)(t, \theta, y) = \int_{\mathbb{R}^{d_1}} f(x, y)\delta(t - g(x, \theta))dx. \tag{6}$$

When $g(x,\theta) = \langle x, \theta \rangle$, PGRT reverts into Partial Radon Transform (PRT) [34].

Proposition 1. *For some defining function g such as linear, circular, and homogeneous polynomials with an odd degree; the Partial Generalized Radon Transform is injective, i.e., for any functions $f_1, f_2 \in \mathbb{L}^1(\mathbb{R}^d)$, $(\mathcal{PGR}f_1)(t,\theta,y) = (\mathcal{PGR}f_2)(t,\theta,y)\ \forall t,\theta,y$ implies $f_1 = f_2$.*

The proof of Proposition 1 is given in Appendix A.1. The main idea to prove the injectivity of PGRT is to show that given a fixed y, the PGRT is the GRT of $f(\cdot, y)$.

Definition 2 (Hierarchical Hybrid Radon Transform)**.** *Given defining functions $g_1 : \mathbb{R}^{d_1} \times \Omega_1 \to \mathbb{R}$ and $g_2 : \mathbb{R}^{d_1} \times \Omega_2 \to \mathbb{R}$, Hierarchical Hybrid Radon Transform $\mathcal{HHR} : \mathbb{L}_1(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}) \to \mathbb{L}_1(\mathbb{R} \times \Omega_1 \times \Omega_2 \times \mathbb{S})$ is defined as:*

$$(\mathcal{HHR}f)(t,\theta_1,\theta_2,\psi) = \int_{\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}} f(x_1,x_2)\delta\left(t - \psi_1 g_1(x_1,\theta_1) - \psi_2 g_2(x_2,\theta_2)\right) dx_1 dx_2, \tag{7}$$

where $\psi = (\psi_1, \psi_2) \in \mathbb{S}$.

The reason for using PRT for the final transform is that the previous PGRTs are assumed to be able to transform the non-linear structure to a linear line. However, PGRT can still be used as a replacement for PRT. Definition 2 can be extended to more than two marginals (see Appendix B).

Proposition 2. *For some defining functions g_1, g_2 such as linear, circular, and homogeneous polynomials with an odd degree; Hierarchical Hybrid Radon Transform is injective, i.e., for any functions $f_1, f_2 \in \mathbb{L}_1(\mathbb{R}^d)$, $(\mathcal{HHR}f_1)(t,\theta_1,\theta_2,\psi) = (\mathcal{HHR}f_2)(t,\theta_1,\theta_2,\psi)\ \forall t,\theta_1,\theta_2,\psi$ implies $f_1 = f_2$.*

The proof of Proposition 2 is given in Appendix A.2. The main idea to prove the injectivity of HHRT is to show that HHRT is the composition of PRT and multiple PGRTs.

HHRT of discrete distributions. We are given $f(x) = \sum_{i=1}^n \alpha_i \delta((x_1,x_2) - (x_{1i},x_{2i}))$ ($n \geq 1$, $\alpha_i \geq 0\ \forall i$). The HHRT of $f(x)$ is $(\mathcal{HHR}f)(t,\theta_1,\theta_2,\psi) = \sum_{i=1}^n \alpha_i \delta\left(t - \psi_1 g_1(x_{i1},\theta_1) - \psi_2 g_2(x_{i2},\theta_2)\right)$. For g_1 and g_2 that are the linear function and (or) the circular function, the time complexity of the transform is $\mathcal{O}(d_1 + d_2)$ which is the same as the complexity of using RT and GRT directly. However, HHRT has an additional constant complexity scaling linearly with the number of marginals, i.e., two marginals in Definition 2.

Example 1. *In this paper, we focus on 3D shape data (mesh) with points and normals representation, i.e., shapes as points representation [47]. In particular, we can transform a 3D shape into a set of points and normals by sampling from the surface of the mesh. In addition, we can convert back to the 3D shape from points and normals with Poisson surface reconstruction [26] algorithm. In this setup, a shape is represented by a 6-dimensional vector $x = (x_1, x_2)$ where $x_1 \in \mathcal{X}_1 \in \mathbb{R}^3$ and $x_2 \in \mathcal{X}_2 \in \mathbb{S}^2$. For the set $\mathcal{X}_1 \in \mathbb{R}^3$, we can use directly the linear defining function $g_1(x_1,\theta_1) = \langle x_1, \theta_1 \rangle$ with $\theta_1 \in \mathbb{S}^2$. For the set $\mathcal{X}_2 \in \mathbb{S}^2$, we can utilize the circular defining function $g_2(x_2,\theta_2) = \|x_2 - r\theta_2\|_2$ with $r \in \mathbb{R}^+$ and $\theta_2 \in \mathbb{S}^2$. As alternative options for $\mathcal{X}_2$, we can use other defining functions from special cases of GRT including Vertical Slice Transform [50], Parallel Slice Transform [51], Spherical Radon Transform [6], and Stereographic Spherical Radon Transform [55].*

Inversion. In Proposition 2, we show that HHRT is the composition of PRT and multiple PGRTs. Therefore, the inversion of HHRT is the composition of the inversion of multiple PGRT (invertibility of PGRT depends on the choice of defining functions [3, 29]) and the inversion of PRT [24].

3.2 Hierarchical Hybrid Sliced Wasserstein Distance

By using HHRT, we obtain a novel variant of SW which is specifically designed for comparing heterogeneous joint distributions.

Definitions. We now define the Hierarchical Hybrid Sliced Wasserstein (H2SW) distance.

Definition 3. *For $p \geq 1$, defining functions g_1, g_2, the hierarchical hybrid sliced Wasserstein-p (H2SW) distance between two distributions $\mu \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$ and $\nu \in \mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_2)$ with an one-dimensional ground metric $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is defined as:*

$$H2SW_p^p(\mu, \nu; c, g_1, g_2) = \mathbb{E}_{(\theta_1,\theta_2,\psi)\sim\mathcal{U}(\Omega_1\times\Omega_2\times\mathbb{S})}[\, W_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu, \mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\nu; c)], \tag{8}$$

where $\mathcal{HHR}_{\theta_1,\theta_2,\psi}\sharp\mu$ and $\mathcal{HHR}_{\theta_1,\theta_2,\psi}\sharp\nu$ are the one-dimensional push-forward distributions created by applying HHRT.

Definition 3 can be easily extended to more than two marginals (see Appendix B)

Topological Properties. We first show that H2SW is a valid metric on the space of distributions on any sets $\mathcal{X} \times \mathcal{Y} \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ $(d_1, d_2 \geq 1)$.

Theorem 1. *For any $p \geq 1$, ground metric c, and defining functions g_1, g_2 which lead to the injectivity of GRT, the hierarchical hybrid sliced Wasserstein $H2SW_p(\cdot, \cdot; c, g_1, g_2)$ is a metric on $\mathcal{P}(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})$ i.e., it satisfies the symmetry, non-negativity, triangle inequality, and identity of indiscernible.*

The proof of Theorem 1 is given in Appendix A.3. It is worth noting that the identity of indiscernible property is proved by the injectivity of HHRT (Proposition 2). We now discuss the connection of H2SW to GSW and Wasserstein distance in some specific cases.

Proposition 3. *For any $p \geq 1$, $c(x, y) = |x - y|$, and $\mu, \nu \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})$, we have:*
(i) $H2SW_p(\mu, \nu; c, g_1, g_2) \leq GSW_p(\mu_1, \nu_1; c, g_1) + GSW_p(\mu_2, \nu_2; c, g_2)$, where $\mu_1(X) = \mu(X \times \mathbb{R}^{d_2})$ and $\mu_2(Y) = \mu(\mathbb{R}^{d_1} \times Y)$ (similar with ν_1 and ν_2).

(ii) If g_1, g_2 are linear defining functions, $H2SW_p(\mu, \nu; c, g_1, g_2) \leq W_p(\mu_1, \nu_1; c) + W_p(\mu_2, \nu_2; c)$.

(iii) If $p = 1$, g_1, g_2 are linear defining functions, $H2SW_1(\mu, \nu; c, g_1, g_2) \leq W_1(\mu, \nu; c, c)$.

The proof of Proposition 3 is given in Appendix A.4.

Sample Complexity. We now discuss the sample complexity of H2SW.

Proposition 4. *For any $p \geq 1$, dimension $d_1, d_2 \geq 1$, $q > p$, $c(x, y) = |x - y|$, g_1, g_2 are linear defining functions or circular defining functions, and $\mu, \nu \in \mathcal{P}_q(\mathbb{R}_1^d \times \mathbb{R}^{d_2})$ with the corresponding distributions μ_n and ν_n $(n \geq 1)$, there exists a constant $C_{p,q}$ depending on p, q such that:*

$$\begin{aligned}
&\mathbb{E}\,|H2SW_p(\mu_n, \nu_n; c, g_1, g_2) - H2SW_p(\mu, \nu; c, g_1, g_2)| \\
&\leq C_{p,q}^{\frac{1}{p}} \left(\sum_{i=0}^{q} q^i C^{q-i}_{g_1,g_2} (M_i(\mu) + M_i(\nu)) \right)^{\frac{1}{p}} \begin{cases} n^{-1/2p} \text{ if } q > 2p, \\ n^{-1/2p} \log(n)^{\frac{1}{p}} \text{ if } q = 2p, \\ n^{-(q-p)/pq} \text{ if } q \in (p, 2p), \end{cases}
\end{aligned} \tag{9}$$

where $M_q(\mu)$ and $M_q(\nu)$ are the q-th moments of μ and ν, C_{g_1,g_2} is a constant depends on g_1, g_2.

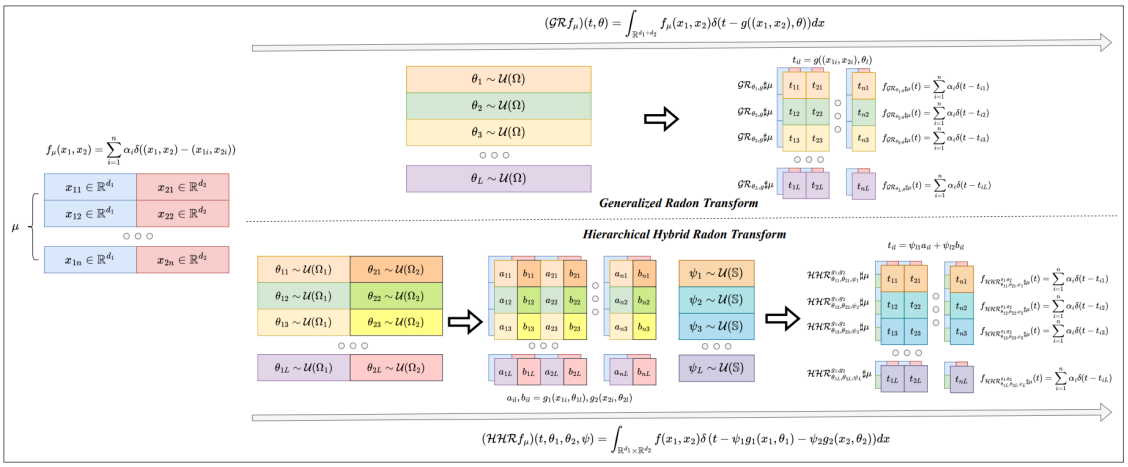

Figure 1: Generalized Radon Transform and Hierarchical Hybrid Radon Transform on a discrete distribution.

The proof of Proposition 4 is given in Appendix A.5. The rate in Proposition 4 is as good as the rate of SW in [39], however, it is slightly worse than the rate of SW in [45, 37, 5] due to the usage of the circular defining functions and simpler assumptions. To the best of our knowledge, the sample complexity of GSW has not been investigated.

Monte Carlo Estimation. Since the expectation in H2SW (Equation 8) is intractable, Monte Carlo estimation and Quasi-Monte Carlo approximation [40] can be used to form a practical evaluation of H2SW. Here, we utilize Monte Carlo estimation for simplicity. In particular, we sample $\theta_{11}, \ldots, \theta_{1L} \overset{i.i.d}{\sim} \mathcal{U}(\Omega_1)$, $\theta_{21}, \ldots, \theta_{2L} \overset{i.i.d}{\sim} \mathcal{U}(\Omega_2)$, and $\psi_1, \ldots, \psi_L \overset{i.i.d}{\sim} \mathcal{U}(\mathbb{S})$. After that, we form the following estimation of H2SW:

$$\widehat{\text{H2SW}}_p^p(\mu, \nu; c, g_1, g_2, L) = \frac{1}{L} \sum_{l=1}^{L} \text{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_{1l},\theta_{2l},\psi_l} \sharp \mu, \mathcal{HHR}^{g_1,g_2}_{\theta_{1l},\theta_{2l},\psi_l} \sharp \nu; c). \tag{10}$$

Proposition 5. *For any $p \geq 1$, dimension $d_1, d_2 \geq 1$, and $\mu, \nu \in \mathcal{P}(\mathbb{R}_1^d \times \mathbb{R}^{d_2})$, we have:*

$$\begin{aligned} &\mathbb{E}|\widehat{H2SW}_p^p(\mu, \nu; c, g_1, g_2, L) - H2SW_p^p(\mu, \nu; c, g_1, g_2)| \\ &\leq \frac{1}{\sqrt{L}} Var\left[W_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi} \sharp \mu, \mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi} \sharp \nu; c) \right]^{\frac{1}{2}}, \end{aligned} \tag{11}$$

where the variance is with respect to $\mathcal{U}(\Omega_1 \times \Omega_2 \times \mathbb{S})$.

The proof of Proposition 5 is given in Appendix A.6. From the proposition, we see that the estimation error of H2SW is the same as SW which is $\mathcal{O}(L^{-1/2})$.

Computational Complexities. The time complexity and memory complexity of H2SW with linear and circular defining functions are $\mathcal{O}(Ln \log n + L(d_1 + d_2 + k)n)$ and $\mathcal{O}(Ln + (d_1 + d_2 + k)n)$ with k is the number of marginals i.e., 2. We can see that the complexities of H2SW are the same as those of SW in terms of the number of supports n and the number of dimensions d. We demonstrate the process of HHRT compared to GRT on a discrete distribution with L realization of θ_1, θ_2, ψ in Figure 1. Overall, the complexities of defining functions are often different in the number of dimensions, hence, H2SW is always scaled the same as SW in the number of supports i.e., $\mathcal{O}(n \log n)$.

Gradient Estimation. In applications, it is desirable to estimate the gradient $\nabla_\phi \text{H2SW}_p^p(\mu_\phi, \nu; c, g_1, g_2)$. We can move the gradient operator to inside the expectation and then apply Monte Carlo estimation. The gradient $\nabla_\phi \text{W}_p^p(\mathcal{HHR}_{\theta_1,\theta_2,\psi}^{g_1,g_2}\sharp\mu_\phi, \mathcal{HHR}_{\theta_1,\theta_2,\psi}^{g_1,g_2}\sharp\nu; c)]$ can be computed easily since the functions g_1, g_2 are usually differentiable.

Beyond uniform slicing distribution. H2SW is defined with the uniform slicing distribution in Definition 3, however, it is possible to extend it to other slicing distributions such as the maximal projecting direction [18], distributional slicing distribution [43], and energy-based slicing distribution [42]. Since the choice of slicing distribution is independent of the main contribution i.e., the slicing operator, we leave this investigation to future work.

H2SW for distributions on the product of Hadamard manifolds. A recent work [9] extends sliced Wasserstein on hyperbolic manifolds [7] and on the manifold of symmetric positive definite matrices [10] to Hadamard manifolds i.e., manifold non-positive curvature. The work discusses the extension of SW to the product of Hadamard manifolds. For the geodesic projection, the closed-form for the projection is intractable. For the Buseman projection, the Buseman projection on the product manifolds is the weighted sum of the Buseman projection with the weights belonging to the unit-sphere. In the work, the weights are a fixed hyperparameter i.e., Cartan-Hadamard Sliced-Wasserstein (CHSW) utilizes only one Buseman function to project the joint distribution. In contrast, H2SW utilizes the Radon Transform on the joint spaces of projections i.e., considering all distributed weighted combinations which is equivalent to considering all Buseman functions under a probability law. As a result, the H2SW is a valid metric as long as the Buseman projections can be proven to be injective (the injectivity of the Buseman projection has not been known at the moment) while Cartan-Hadamard Sliced-Wasserstein is only pseudo metric since the injectivity of a fixed weighted combination is not trivial to show. Moreover, H2SW does not only focus on the product of Hadamard manifolds i.e., H2SW is a generic distance for heterogeneous joint distributions in which marginal domains do not necessary manifolds e.g., images [41], functions [22], and so on. In the later experiments, we conduct experiments on comparing 3D shapes which are represented by a distribution on the product of the Euclidean space and the 2D sphere (not a Hadamard manifold).

4 Experiments

In this section, we first compare the performance of the proposed H2SW with SW and GSW in the 3D mesh deformation application. After that, we further evaluate the performance of H2SW in training a deep 3D mesh autoencoder compared to SW and GSW. Finally, we compare H2SW with SW and Cartan-Hadamard Sliced-Wasserstein (CHSW) in comparing datasets on the product of Hadamard manifolds. In the experiments, we use $c(x, y) = |x - y|$ and $p = 2$ for all SW variants.

4.1 3D Mesh Deformation

In this task, we would like to move from a source mesh to a target mesh. To represent those meshes, we sample 10000 points by Poisson disk sampling and their corresponding normal vectors of the mesh surface at those points. Let the source mesh be denoted as $X(0) = \{x_1(0), \ldots, x_n(0)\}$ and the target mesh be denoted as $Y = \{y_1, \ldots, y_n\}$. We deform $X(0)$ to Y by integrating the ordinary differential equation $\dot{X}(t) = -n\nabla_{X(t)}\left[\mathcal{S}\left(\frac{1}{n}\sum_{i=1}^n \delta(x - x_i(t)), \frac{1}{n}\sum_{i=1}^n \delta(y - y_i)\right)\right]$, where $\mathcal{S}$ denotes a SW variant. We utilize the Euler discretization scheme with step size 0.01 and 5000 steps. For evaluation, we use the joint Wasserstein distance in Equation 5 with the mixed distance from the

Table 1: Summary of joint Wasserstein distances (multiplied by 100) across time steps from deformation from the sphere mesh to the Stanford Bunny mesh.

Distances	Step 100 ($W_{c_1,c_2} \downarrow$)	Step 300 ($W_{c_1,c_2} \downarrow$)	Step 600 ($W_{c_1,c_2} \downarrow$)	Step 3000 ($W_{c_1,c_2} \downarrow$)	Step 4000 ($W_{c_1,c_2} \downarrow$)	Step 5000 ($W_{c_1,c_2} \downarrow$)
SW L=10	27.359±0.197	4.518±0.074	1.411±0.051	0.691±0.057	0.218±0.035	0.162±0.02
GSW L=10	26.956±0.13	4.356±0.077	1.453±0.056	0.66±0.012	0.206±0.028	0.152±0.026
H2SW L=10	**24.109±0.055**	**2.568±0.055**	**1.168±0.106**	**0.543±0.081**	**0.093±0.023**	**0.06±0.007**
SW L=100	26.903±0.028	4.161±0.052	1.195±0.074	0.439±0.024	0.069±0.018	0.059±0.015
GSW L=100	26.63±0.094	4.021±0.093	1.234±0.08	**0.429±0.016**	0.072±0.013	0.044±0.009
H2SW L=100	**23.85±0.067**	**2.433±0.088**	**1.173±0.079**	0.452±0.042	**0.063±0.008**	**0.029±0.006**

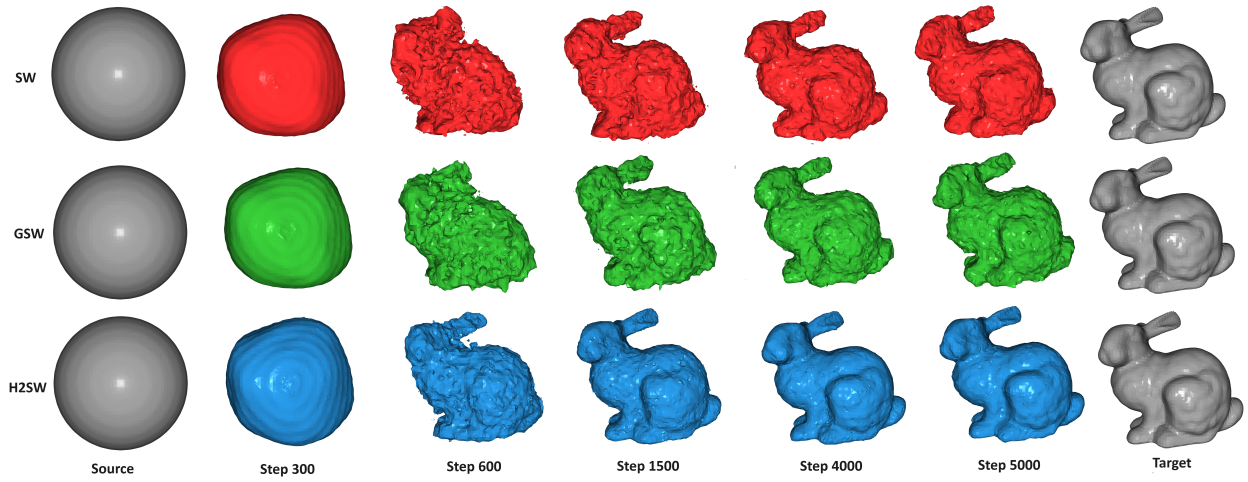

Figure 2: Visualization of deformation from the sphere mesh to the Stanford Bunny mesh with $L = 10$.

Euclidean distance and the great circle distance. We use the circular defining function for GSW, and use the linear defining function and the circular defining function for H2SW. We vary the number of projections $L \in \{10, 100\}$ for all variants. For H2SW and GSW, we select the best hyperparameter of the circular defining function $r \in \{0.5, 0.7, 0.8, 0.9, 1, 5, 10, 50, 100\}$.

Results. We compare H2SW with GSW and SW by deforming the sphere mesh to the Stanford Bunny mesh [16, 61]. We report the quantitative results in Table 1 after 3 independent runs and the qualitative result for $L = 10$ in Figure 2 and $L = 100$ in Figure 5 in Appendix D. From Table 1, we observe that H2SW helps the deformation convergence faster and better in terms of the joint Wasserstein distance, especially for a small value of the number of projections i.e., $L = 10$. The result for $L = 100$ is better than $L = 10$ which is consistent with Proposition 5. The qualitative results in Figure 2 and Figure 5 also reinforce the favorable performance of H2SW since they are visually consistent with quantitative scores. We also conduct deformation to the Armadillo mesh [61] in Table 4, Figure 6, and Figure 7 in Appendix D and we observe the same phenomenon that H2SW is the best variant for 3D meshes. From those experiments, H2SW has shown the benefit of the HHRT in transforming a joint distribution over the product of the Euclidean space and the 2D sphere compared to the conventional RT of SW and the conventional GRT of GSW.

Table 2: Joint Wasserstein distance reconstruction errors (multiplied by 100) from three different runs of autoencoders trained by SW, GSW, and H2SW with the number of projections $L = 100$ and $L = 1000$.

Distance	Epoch 500 $\text{W}_{c_1,c_2}(\downarrow)$	Epoch 1000 $\text{W}_{c_1,c_2}(\downarrow)$	Epoch 2000 $\text{W}_{c_1,c_2}(\downarrow)$
SW L=100	136.87 ± 1.18	133.24 ± 0.50	131.10 ± 0.34
GSW L=100	136.51 ± 0.20	133.36 ± 0.24	130.80 ± 0.46
H2SW L=100	$\mathbf{135.54 \pm 0.72}$	$\mathbf{132.24 \pm 0.36}$	$\mathbf{130.24 \pm 0.47}$
SW L=1000	135.85 ± 0.92	132.88 ± 0.27	130.93 ± 0.11
GSW L=1000	136.40 ± 0.10	133.02 ± 0.98	130.76 ± 0.26
H2SW L=1000	$\mathbf{135.47 \pm 0.64}$	$\mathbf{132.17 \pm 0.10}$	$\mathbf{129.87 \pm 0.44}$

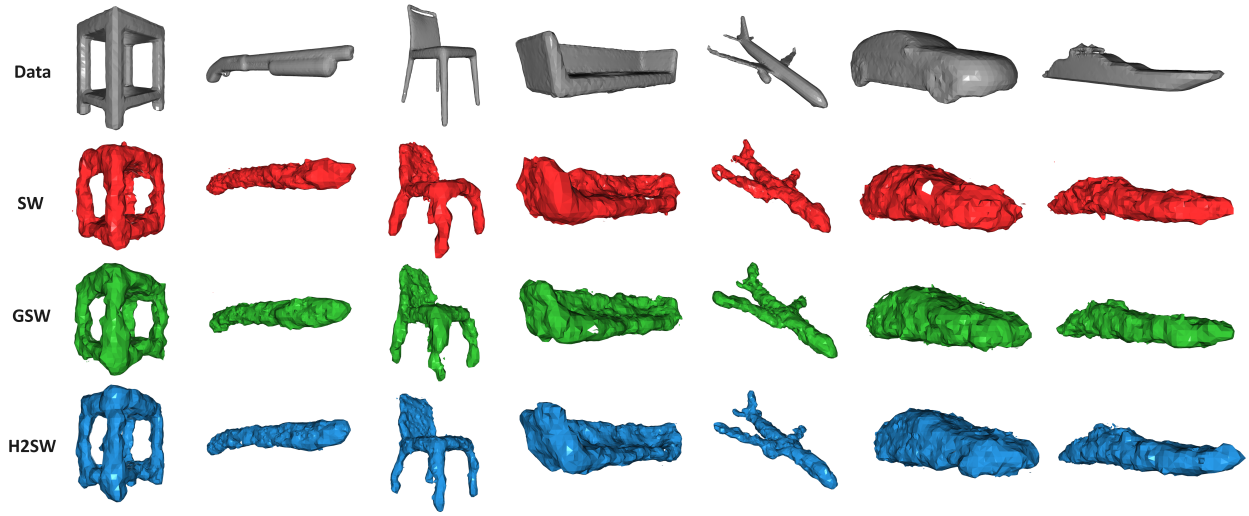

Figure 3: Visualization of some randomly selected reconstruction meshes from autoencoders trained by SW, GSW, and H2SW in turn with the number of projections $L = 100$ at epoch 2000.

4.2 Training deep 3D mesh autoencoder

We utilize the processed ShapeNet dataset [12] from [47], then sample 2048 points and the corresponding normal vectors from each shape in the dataset. Formally, we would like to train an autoencoder that contains an encoder f_ϕ that maps a mesh $X \in \mathbb{R}^{2048\times 6}$ to a latent code $z \in \mathbb{R}^{1024}$, and a decoder g_ψ that maps the latent code z back to the reconstructed mesh $\tilde{X} \in \mathbb{R}^{2048\times 6}$. We adopt Point-Net [49] architecture to construct the autoencoder. We want to train the encoder f_ϕ and the decoder g_ψ such that $\tilde{X} = g_\psi(f_\phi(X)) \approx X$ for all shapes X in the dataset. To do that, we solve the following optimization problem:

$$\min_{\phi,\gamma} \mathbb{E}_{X\sim\mu(X)}[\mathcal{S}(P_X, P_{g_\gamma(f_\phi(X))})],$$

where $\mathcal{S}$ is a sliced Wasserstein variant, and $P_X = \frac{1}{n}\sum_{i=1}^n \delta(x - x_i)$ denotes the empirical distribution over the point cloud $X = (x_1, \ldots, x_n)$. We train the autoencoder for 2000 epochs on the training set of the ShapeNet dataset using an SGD optimizer with a learning rate of $1e-3$, and a batch size of 128. For evaluation, we also use the joint Wasserstein distance in Equation 5 with the mixed distance from the Euclidean distance and the great circle distance to measure the average reconstruction loss

Table 3: Relative error to the joint Wasserstein distance of SW, CHSW, and H2SW.

Distances	$L = 100$	$L = 500$	$L = 1000$	$L = 2000$
SW	4.618 ± 0.744	4.253 ± 0.398	4.235 ± 0.310	4.198 ± 0.238
CHSW	4.449 ± 0.497	4.063 ± 0.254	4.059 ± 0.167	4.035 ± 0.145
H2SW	$\mathbf{4.381 \pm 0.695}$	$\mathbf{4.001 \pm 0.267}$	$\mathbf{4.048 \pm 0.182}$	$\mathbf{3.998 \pm 0.142}$

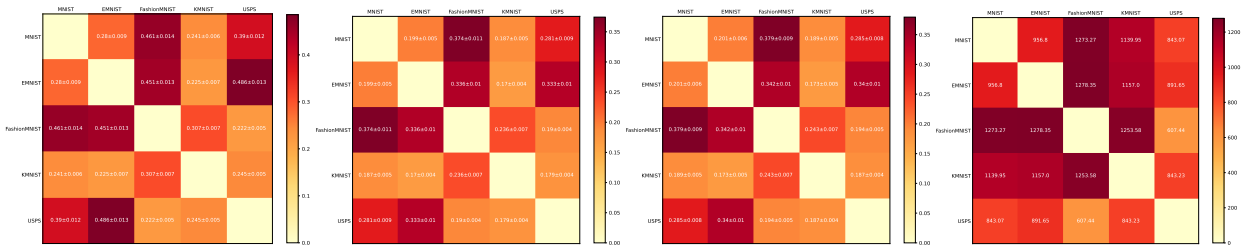

Figure 4: Cost matrices between datasets from SW, CHSW, and H2SW with $L = 2000$.

on the testing set of the ShapeNet dataset. We use the circular defining function for GSW, and use the linear defining function and the circular defining function for H2SW. For H2SW and GSW, we select the best hyperparameter of the circular defining function $r \in \{0.5, 0.7, 0.8, 0.9, 1, 5, 10\}$. For more details such as the neural network architectures, we refer the reader to Appendix B.

Results. We report the joint Wasserstein reconstruction errors (measured in three independent times) on the testing set in Table 2 with trained autoencoder at epoch 500, 1000, and 2000 from SW, GSW, and H2SW with the number of projections $L = 100$ and $L = 1000$. In addition, we show some randomly reconstructed meshes for epoch 2000in Figure 3 and for epoch 500 in Figure 8 in Appendix D. From Table 2, we observe that H2SW yields the lowest reconstruction errors for both $L = 100$ and $L = 1000$. Moreover, we see that the reconstruction errors are lower with $L = 1000$ than ones with $L = 100$ for all SW variants. The qualitative reconstructed meshes in Figure 3 and Figure 8reflect the same relative comparison. It is worth noting that both the qualitative and the qualitative performance of autoencoders can be improved by using more powerful neural networks. Since we focus on comparing SW, GSW, and H2SW, we only use a light neural network i.e., Point-Net [49] architecture. The trained autoencoders can be further used to reduce the size of 3D meshes for data compression and for dimension reduction, however, such downstream applications are not our focus in the current investigation of the paper.

4.3 Comparing Datasets on The Product of Hadamard Manifolds

We follow the same experimental setting from [9]. Here, we have datasets as sets of feature-label pairs which are embedded in the space of $\mathbb{R}^{d_1} \times \mathbb{L}^{d_2}$ where $\mathbb{L}^{d_2}$ denotes a Lorentz model of d_2 dimension (a hyperbolic space). We uses MNIST [32] dataset, EMNIST dataset [14], Fashion MNIST dataset [58], KMNIST dataset [13], and USPS dataset [25]. For CHSW, we use Buseman projection on the product space of Euclidean and the Lorentz model. For H2SW, we use the linear defining function and the Buseman function on the Lorentz model. We refer the reader to Appendix B for greater detail on Buseman functions and experimental setups. We compare SW, CHSW, and H2SW by

varying $L \in \{100, 500, 1000, 2000\}$ For evaluation, we use the joint Wasserstein distance in [1] as the ground truth. In particular, let C_W be the cost matrix from the joint Wasserstein distance and C be a given cost matrix, we use $|C/\max(C) - C_W/\max(C_W)|$ as the relative error.

Results. We report the relative errors from SW, CHSW, and H2SW in Table 3 after 100 independent runs. In addition, we show the cost matrices from SW, CHSW, H2SW. and joint Wasserstein distance with $L = 2000$ in Figure 4. Cost matrices for $L = 100$, $L = 500$, and $L = 1000$ are given in Figure 9-11 in Appendix D. From Table 3, we see that H2SW gives a lower relative error than CHSW and SW. Therefore, using H2SW for comparing datasets is the most equivalent to the joint Wasserstein distance in terms of the relative error. We also observe that increasing the value of the number of projections also reduces the relative errors for all SW variants. Again, we would like to recall that H2SW can be used for heterogeneous joint distributions beyond the product of Hadamard manifolds as shown in previous experiments.

5 Conclusion

We have presented Hierarchical Hybrid Sliced Wasserstein (H2SW) distance, a novel sliced probability metric for heterogeneous joint distributions i.e., joint distributions have marginals on different domains. The key component of H2SW is the proposed hierarchical hybrid Radon Transform (HHRT) which is the composition of partial Radon Transform and multiples proposed partial generalized Radon Transform. We then discuss the injectivity of the proposed transforms and theoretical properties of H2SW including topological properties, statistical properties, and computational properties. On the experimental side, we show that H2SW has favorable performance in applications of 3D mesh deformation, training deep 3D mesh autoencoder, and datasets comparison. In those applications, heterogeneous joint distributions appear in the form of joint distributions on the product of Euclidean space and 2D sphere, and the product of Hadamard manifolds. In the future, we will extend the application of H2SW to more complicated heterogeneous joint distributions.

Supplement to "Hierarchical Hybrid Sliced Wasserstein: A Scalable Metric for Heterogeneous Joint Distributions"

We first provide skipped proofs in the main paper in Appendix A. We when provide some additional materials including additional background and extended definitions om Appendix B. After that, we discuss some related works in Appendix C. We report additional experimental results in Appendix D. Finally, we report computational infrastructure in Appendix E.

A Proofs

In this appendix, we provide proofs for main results in the paper.

A.1 Proof of Proposition 1

For any t, θ, y, we are given $(\mathcal{PGR}f_1)(t, \theta, y) = (\mathcal{PGR}f_2)(t, \theta, y)$. By Definition 1, we obtain that

$$\int_{\mathbb{R}^{d_1}} f_1(x, y)\delta(t - g(x, \theta))dx = \int_{\mathbb{R}^{d_1}} f_2(x, y)\delta(t - g(x, \theta))dx.$$

For any $\varepsilon \in \mathbb{R}^2$, we find that

$$\int_{\mathbb{R}^{d_2}} \int_{\mathbb{R}^{d_1}} f_1(x, y)\delta(t - g(x, \theta))e^{-i2\pi\langle\varepsilon,y\rangle}dxdy = \int_{\mathbb{R}^{d_2}} \int_{\mathbb{R}^{d_1}} f_2(x, y)\delta(t - g(x, \theta))e^{-i2\pi\langle\varepsilon,y\rangle}dxdy.$$

Applying the Fubini's theorem, we have

$$\int_{\mathbb{R}^{d_1}} f_1(x, y) \int_{\mathbb{R}^{d_2}} e^{-i2\pi\langle\varepsilon,y\rangle}dy\delta(t - g(x, \theta))dx = \int_{\mathbb{R}^{d_1}} \int_{\mathbb{R}^{d_2}} f_2(x, y)e^{-i2\pi\langle\varepsilon,y\rangle}dy\delta(t - g(x, \theta))dx,$$

which leads to the following equation:

$$\left(\mathcal{GR}\int_{\mathbb{R}^{d_2}} f_1(x, y)e^{-i2\pi\langle\varepsilon,y\rangle}dy\right) = \left(\mathcal{GR}\int_{\mathbb{R}^{d_2}} f_2(x, y)e^{-i2\pi\langle\varepsilon,y\rangle}dy\right).$$

By the injectivity of GRT, we have

$$\int_{\mathbb{R}^{d_2}} f_1(x, y)e^{-i2\pi\langle\varepsilon,y\rangle}dy = \int_{\mathbb{R}^{d_2}} f_2(x, y)e^{-i2\pi\langle\varepsilon,y\rangle}dy.$$

Then, for any $\epsilon \in \mathbb{R}^{d_1}$, we arrive a the following equation:

$$\int_{\mathbb{R}^{d_1}} \int_{\mathbb{R}^{d_2}} f_1(x, y)e^{-i2\pi\langle\varepsilon,y\rangle}e^{-i2\pi\langle\epsilon,x\rangle}dydx = \int_{\mathbb{R}^{d_1}} \int_{\mathbb{R}^{d_2}} f_2(x, y)e^{-i2\pi\langle\varepsilon,y\rangle}e^{-i2\pi\langle\epsilon,x\rangle}dydx.$$

which is $(\mathcal{F}f_1(x, y)) = (\mathcal{F}f_2(x, y)))$ with $\mathcal{F}$ denotes the Fourier transform. By the injectivity of the Fourier Transform, we have $f_1(x, y) = f_2(x, y)$ for any x, y, which concludes the proof.

A.2 Proof of Proposition 2

We first show that HHRT is the composition of PGRT and PRT. We have

$$
\begin{aligned}
&(\mathcal{PR}(\mathcal{PGR}(\mathcal{PGR}f)))(t,\theta_1,\theta_2,\psi)\\
&=\int_{\mathbb{R}^2}\int_{\mathbb{R}^{d_1}}\int_{\mathbb{R}^{d_2}} f(x,y)\delta(t_1-g_1(x,\theta_1))\delta(t_2-g_2(y,\theta_2))\delta(t-\psi_1 t_1-\psi_2 t_2)dxdydt_1dt_2\\
&=\int_{\mathbb{R}^{d_1}}\int_{\mathbb{R}^{d_2}} f(x,y)\int_{\mathbb{R}^2}\delta(t_1-g_1(x,\theta_1))\delta(t_2-g_2(y,\theta_2))\delta(t-\psi_1 t_1-\psi_2 t_2)dt_1dt_2dxdy\\
&=\int_{\mathbb{R}^{d_1}}\int_{\mathbb{R}^{d_2}} f(x,y)\delta\left(t-\psi_1 g_1(x,\theta_1)-\psi_2 g_2(y,\theta_2)\right)dxdy\\
&=(\mathcal{HHR}f)(t,\theta_1,\theta_2,\psi).
\end{aligned}
$$

For any t,θ_1,θ_2,ψ, we are given $(\mathcal{HHR}f_1)(t,\theta_1,\theta_2,\psi)=(\mathcal{HHR}f_2)(t,\theta_1,\theta_2,\psi)$, which is equivalent to:

$$
(\mathcal{PR}(\mathcal{PGR}(\mathcal{PGR}f_1)))(t,\theta_1,\theta_2,\psi)=(\mathcal{PR}(\mathcal{PGR}(\mathcal{PGR}f_2)))(t,\theta_1,\theta_2,\psi).
$$

By the injectivity of the PRT and the PGRT, we obtain $f_1(x,y)=f_2(x,y)$ for any x,y which completes the proof.

A.3 Proof of Theorem 1

To prove that the hierarchical hybrid sliced Wasserstein $H2SW_p(\cdot,\cdot;c,g_1,g_2)$ is a metric on the space of distributions on $\mathcal{P}(\mathbb{R}^{d_1}\times\mathbb{R}^{d_2})$ for any $p\geq 1$, ground metric c, and defining functions g_1,g_2, we need to show that it satisfies non-negativity, symmetry, triangle inequality, and identity of indiscernible.

Non-Negativity. Since $\mathrm{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu,\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\nu;c)\geq 0$ [48] for any θ_1,θ_2,ψ, we have:

$$
\mathbb{E}_{(\theta_1,\theta_2,\psi)\sim\mathcal{U}(\Omega_1\times\Omega_2\times\mathbb{S})}[\mathrm{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu,\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\nu;c)]\geq 0,
$$

which means that $H2SW_p(\mu,\nu;c,g_1,g_2)\geq 0$ for any μ and ν.

Symmetry. Since we have the symmetry of the Wasserstein distance $\mathrm{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu,\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\nu;c)=\mathrm{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\nu,\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu;c)$ [48] for any θ_1,θ_2,ψ, we have:

$$
\begin{aligned}
&\mathbb{E}_{(\theta_1,\theta_2,\psi)\sim\mathcal{U}(\Omega_1\times\Omega_2\times\mathbb{S})}[\mathrm{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu,\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\nu;c)]\\
&=\mathbb{E}_{(\theta_1,\theta_2,\psi)\sim\mathcal{U}(\Omega_1\times\Omega_2\times\mathbb{S})}[\mathrm{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\nu,\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu;c)],
\end{aligned}
$$

which means that $H2SW_p(\mu,\nu;c,g_1,g_2)=H2SW_p(\nu,\mu;c,g_1,g_2)$ any μ and ν.

Triangle Inequality. Given c to be a valid metric on $\mathbb{R}$, we can use the triangle inequality of the

Wasserstein distance. For any distributions $\mu_1, \mu_2, \mu_3 \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})$, we find that

$$\begin{aligned}
\text{H2SW}_p(\mu_1, \mu_2; c, g_1, g_2) &= \left(\mathbb{E}_{(\theta_1,\theta_2,\psi)\sim\mathcal{U}(\Omega_1\times\Omega_2\times\mathbb{S})}[\text{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_1, \mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_2; c)]\right)^{\frac{1}{p}} \\
&\leq \Big(\mathbb{E}_{(\theta_1,\theta_2,\psi)\sim\mathcal{U}(\Omega_1\times\Omega_2\times\mathbb{S})}[(\text{W}_p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_1, \mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_3; c) \\
&\quad + \text{W}_p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_3, \mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_2; c))^p]\Big)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}_{(\theta_1,\theta_2,\psi)\sim\mathcal{U}(\Omega_1\times\Omega_2\times\mathbb{S})}[\text{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_1, \mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_3; c)]\right)^{\frac{1}{p}} \\
&\quad + \left(\mathbb{E}_{(\theta_1,\theta_2,\psi)\sim\mathcal{U}(\Omega_1\times\Omega_2\times\mathbb{S})}[\text{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_3, \mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_2; c)]\right)^{\frac{1}{p}} \\
&= \text{H2SW}_p(\mu_1, \mu_3; c, g_1, g_2) + \text{H2SW}_p(\mu_3, \mu_2; c, g_1, g_2),
\end{aligned}$$

where the final inequality is due to Minkowski's inequality. Therefore, we complete the proof for the triangle inequality of the hierarchical hybrid sliced Wasserstein.

Identity of indiscernible. For any $p \geq 1$, ground metric c, and g_1, g_2, when $\mu = \nu$, we have $\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu = (\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\nu$. Therefore, we have $\text{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_1, \mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_2; c) = 0$ which leads to $\text{H2SW}_p(\mu, \nu; c, g_1, g_2) = 0$. Now, assume that $\text{H2SW}_p(\mu, \nu; c, g_1, g_2) = 0$, then $\text{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_1, \mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu_2; c) = 0$ for almost everywhere $\theta_1 \in \Omega_1, \theta_2 \in \Omega_2, \psi \in \mathbb{S}$. By applying the identity property of the Wasserstein distance, we have $\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu = (\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\nu$ for almost everywhere $\theta_1 \in \Omega_1, \theta_2 \in \Omega_2, \psi \in \mathbb{S}$. Since the HHRT is injective (proved in Proposition 2), we obtain $\mu = \nu$.

A.4 Proof of Proposition 3

(i) For any $p \geq 1$, $c(x, y) = |x - y|$, and $\mu, \nu \in \mathcal{P}(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2})$, we have

$$\begin{aligned}
&\text{H2SW}_p(\mu, \nu; c, g_1, g_2) \\
&= \left(\mathbb{E}_{(\theta_1,\theta_2,\psi)\sim\mathcal{U}(\Omega_1\times\Omega_2\times\mathbb{S})}[\text{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu, \mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\nu; c)]\right)^{\frac{1}{p}} \\
&= \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu,\nu)}\int |\psi_1(g_1(\theta_1, x_1) - g_1(\theta_1, y_1)) + \psi_2(g_2(\theta_2, x_2) - g_1(\theta_2, y_2))|^p d\pi(x_1, x_2, y_1, y_2)\right]\right)^{\frac{1}{p}}
\end{aligned}$$

By applying the Cauchy-Schwartz inequality, we have:

$$\begin{aligned}
&\text{H2SW}_p(\mu, \nu; c, g_1, g_2) \\
&\leq \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu,\nu)}\int (\sqrt{\psi_1^2 + \psi_2^2})^p (\sqrt{(g_1(\theta_1, x_1) - g_1(\theta_1, y_1))^2 + (g_2(\theta_2, x_2) - g_2(\theta_2, y_2))^2})^p d\pi(x_1, x_2, y_1, y_2)\right]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu,\nu)}\int (|g_1(\theta_1, x_1) - g_1(\theta_1, y_1)| + |g_2(\theta_2, x_2) - g_2(\theta_2, y_2)|)^p d\pi(x_1, x_2, y_1, y_2)\right]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu,\nu)}\int |g_1(\theta_1, x_1) - g_1(\theta_1, y_1)|^p d\pi(x_1, x_2, y_1, y_2)\right]\right)^{\frac{1}{p}} \\
&\quad + \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu,\nu)}\int |g_2(\theta_2, x_2) - g_2(\theta_2, y_2)|^p d\pi(x_1, x_2, y_1, y_2)\right]\right)^{\frac{1}{p}}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\nu_1)}\int |g_1(\theta_1,x_1)-g_1(\theta_1,y_1)|^p d\pi(x_1,y_1)\right]\right)^{\frac{1}{p}} \\
&\quad + \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu_2,\nu_2)}\int |g_2(\theta_2,x_2)-g_2(\theta_2,y_2)|^p d\pi(x_2,y_2)\right]\right)^{\frac{1}{p}} \\
&= \mathrm{GSW}_p(\mu_1,\nu_1;g_1,c)+\mathrm{GSW}_p(\mu_2,\nu_2;g_2,c),
\end{aligned}
$$

where the last inequality is due to the Minkowski's inequality.

(ii) From (i), we have $\mathrm{H2SW}_p(\mu,\nu;c,g_1,g_2) \leq \mathrm{GSW}_p(\mu_1,\nu_1;g_1,c)+\mathrm{GSW}_p(\mu_2,\nu_2;g_2,c)$. When, g_1, g_2, and $c(x,y)=|x-y|$ are linear defining functions, we have:

$$
\begin{aligned}
\mathrm{GSW}_p(\mu_1,\nu_1;g_1,c) &= \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\nu_1)}\int (|\theta^\top x_1-\theta^\top y_1|^p d\pi(x_1,y_1)\right]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\nu_1)}\int (\|\theta\|_2\|x_1--y_1\|_2^p d\pi(x_1,y_1)\right]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu_1,\nu_1)}\int (\|x_1-y_1\|^p d\pi(x_1,y_1)\right]\right)^{\frac{1}{p}} \\
&= \left(\inf_{\pi\in\Pi(\mu_1,\nu_1)}\int (\|x_1-y_1\|^p d\pi(x_1,y_1)\right)^{\frac{1}{p}} \\
&= W_p(\mu_1,\nu_1;c).
\end{aligned}
$$

Similarly, we have $\mathrm{GSW}_p(\mu_2,\nu_2;g_1,c) \leq W_p(\mu_2,\nu_2;c)$. Therefore, we obtain the proof of $\mathrm{H2SW}_p(\mu,\nu;c,g_1,g_2) \leq W_p(\mu_1,\nu_1;c)+W_p(\mu_1,\nu_1;c)$.

(iii) When g_1, g_2 are linear defining functions, we have:

$$
\begin{aligned}
\mathrm{H2SW}_p(\mu,\nu;c,g_1,g_2) &\leq \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu,\nu)}\int (|\theta_1^\top x_1-\theta_1^\top y_1)|+|\theta_2^\top x_2-\theta_2^\top y_2|)^p d\pi(x_1,x_2,y_1,y_2)\right]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu,\nu)}\int (|\theta_1^\top x_1-\theta_1^\top y_1)|+|\theta_2^\top x_2-\theta_2^\top y_2|)^p d\pi(x_1,x_2,y_1,y_2)\right]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}\left[\inf_{\pi\in\Pi(\mu,\nu)}\int (|x_1-y_1)|+|x_2-y_2|)^p d\pi(x_1,x_2,y_1,y_2)\right]\right)^{\frac{1}{p}} \\
&= \left(\inf_{\pi\in\Pi(\mu,\nu)}\int (|x_1-y_1)|+|x_2-y_2|)^p d\pi(x_1,x_2,y_1,y_2)\right)^{\frac{1}{p}}
\end{aligned}
$$

When $p=1$, we obtain:

$$
\begin{aligned}
\mathrm{H2SW}_1(\mu,\nu;c,g_1,g_2) &\leq \left(\inf_{\pi\in\Pi(\mu,\nu)}\int (|x_1-y_1)|+|x_2-y_2|) d\pi(x_1,x_2,y_1,y_2)\right)^{\frac{1}{p}} \\
&= W_1(\mu,\nu;c,c),
\end{aligned}
$$

which completes the proof.

A.5 Proof of Proposition 4

Let $p \geq 1$, $c(x,y) = |x-y|$, $\mu \in \mathcal{P}(\mathbb{R})$ with the corresponding empirical distribution μ_n, we assume that there exists $q > p$ such that the $q-$th order moment of μ i.e, $M_q(\mu) = \int_{\mathbb{R}} |x|^q d\mu(x)$, is bounded by $B < \infty$. From Theorem 1 in [21], there exists a constant $C_{p,q}$ such that:

$$\mathbb{E}\left[W_p^p(\mu_n, \mu; c)\right] \leq C_{p,q} B \begin{cases} n^{-1/2} \text{ if } q > 2p, \\ n^{-1/2} \log(n)^{\frac{1}{p}} \text{ if } q = 2p, \\ n^{-(q-p)/q} \text{ if } q \in (p, 2p). \end{cases}$$

We show that $\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi} \sharp \mu$ has finite bounded moments. In particular, we have:

$$\begin{aligned} M_k(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi} \sharp \mu) &= \int_{\mathbb{R}} |t|^k d(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi} \sharp \mu)(t) \\ &= \int_{\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}} |\psi_1 g_1(\theta_1, x_1) + \psi_2 g_2(\theta_2, x_2)|^k d\mu(x_1, x_2) \\ &\leq \int_{\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}} (\psi_1^2 + \psi_2^2)^{k/2} (g_1(\theta_1, x_1)^2 + g_2(\theta_2, x_2)^2)^{k/2} d\mu(x_1, x_2) \\ &\leq \int_{\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}} (|g_1(\theta_1, x_1)| + |g_2(\theta_2, x_2)|)^k d\mu(x_1, x_2), \end{aligned}$$

where the first inequality is due to the Cauchy-Schwarz inequality and the second inequality is due to the fact that $\|x\|_2 \leq |x|$. For the linear defining functions $g(\theta, x) = \theta^\top x$, we have $|g(\theta, x)| = |\theta^\top x| \leq \|x\|_1$. For the circular defining functions $g(\theta, x) = \|x - r\theta\|_2 \leq \|x - r\theta\|_1 \leq \|x\|_1 + \|r\theta\|_1 \leq \|x\|_1 + r$. Therefore, we have:

$$\begin{aligned} M_k(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi} \sharp \mu) &\leq \int_{\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}} (|x_1| + |x_2| + C_{g_1,g_2})^k d\mu(x_1, x_2) \\ &= \int_{\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}} \sum_{i=0}^{k} k^i (|x_1| + |x_2|)^i C_{g_1,g_2}^{k-i} d\mu(x_1, x_2) \\ &= \sum_{i=0}^{k} k^i C_{g_1,g_2}^{k-i} \int_{\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}} (|x_1| + |x_2|)^i d\mu(x_1, x_2) \\ &\leq \sum_{i=0}^{k} k^i C_{g_1,g_2}^{k-i} M_i(\mu), \end{aligned}$$

where $C_{g_1,g_2} = 0$ if g_1, g_2 are linear, $C_{g_1,g_2} = r$ if g_1 and g_2 are linear and circular respectively (exchangeable), and $C_{g_1,g_2} = 2r$ if both g_1 and g_2 are circular.

Now, using the triangle inequality of H2SW (Theorem 1), we have:

$$\begin{aligned} &\mathbb{E}\left|\text{H2SW}_p(\mu_n, \nu_n; c, g_1, g_2) - \text{H2SW}_p(\mu, \nu; c, g_1, g_2)\right| \\ &\quad \leq \mathbb{E}\left|\text{H2SW}_p(\mu, \mu_n; c, g_1, g_2) + \text{H2SW}_p(\nu, \nu_n; c, g_1, g_2)\right| \\ &\quad \leq \mathbb{E}\left|\text{H2SW}_p(\mu, \mu_n; c, g_1, g_2)\right| + \mathbb{E}\left|\text{H2SW}_p(\nu, \nu_n; c, g_1, g_2)\right| \\ &\quad \leq \left(\mathbb{E}\left|\text{H2SW}_p^p(\mu, \mu_n; c, g_1, g_2)\right|\right)^{\frac{1}{p}} + \left(\mathbb{E}\left|\text{H2SW}_p^p(\nu, \nu_n; c, g_1, g_2)\right|\right)^{\frac{1}{p}}, \end{aligned}$$

where the last inequality is due to Holder's inequality. Combining with previous results, we obtain:

$$\begin{aligned}
&\mathbb{E}\left|\mathrm{H2SW}_p(\mu_n,\nu_n;c,g_1,g_2)-\mathrm{H2SW}_p(\mu,\nu;c,g_1,g_2)\right| \\
&\qquad\leq C_{p,q}^{\frac{1}{p}}\left(\sum_{i=0}^{q} q^i C_{g_1,g_2}^{q-i}(M_i(\mu)+M_i(\nu))\right)^{\frac{1}{p}}\begin{cases} n^{-1/2p} \text{ if } q>2p, \\ n^{-1/2p}\log(n)^{\frac{1}{p}} \text{ if } q=2p, \\ n^{-(q-p)/pq} \text{ if } q\in(p,2p), \end{cases}
\end{aligned}$$

which completes the proof.

A.6 Proof of Proposition 5

For any $p\geq 1$, and $\mu,\nu\in\mathcal{P}(\mathbb{R}^{d_1}\times\mathbb{R}^{d_2})$, using the Holder's inequality, we have:

$$\begin{aligned}
&\mathbb{E}|\widehat{\mathrm{H2SW}}_p^p(\mu,\nu;c,g_1,g_2,L)-\mathrm{H2SW}_p^p(\mu,\nu;c,g_1,g_2)| \\
&\leq\left(\mathbb{E}|\widehat{\mathrm{H2SW}}_p^p(\mu,\nu;c,g_1,g_2,L)-\mathrm{H2SW}_p^p(\mu,\nu;c,g_1,g_2)|^2\right)^{\frac{1}{2}} \\
&=\left(\mathbb{E}\left|\frac{1}{L}\sum_{l=1}^{L}\mathrm{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_{1l},\theta_{2l},\psi_l}\sharp\mu,\mathcal{HHR}^{g_1,g_2}_{\theta_{1l},\theta_{2l},\psi_l}\sharp\nu;c)-\mathbb{E}\left[\mathrm{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\mu,\mathcal{HHR}^{g_1,g_2}_{\theta_1,\theta_2,\psi}\sharp\nu;c)\right]\right|^2\right)^{\frac{1}{2}} \\
&=\left(Var\left[\frac{1}{L}\sum_{l=1}^{L}\mathrm{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_{1l},\theta_{2l},\psi_l}\sharp\mu,\mathcal{HHR}^{g_1,g_2}_{\theta_{1l},\theta_{2l},\psi_l}\sharp\nu;c)\right]\right)^{\frac{1}{2}} \\
&=\frac{1}{\sqrt{L}}Var\left[\mathrm{W}_p^p(\mathcal{HHR}^{g_1,g_2}_{\theta_{1l},\theta_{2l},\psi_l}\sharp\mu,\mathcal{HHR}^{g_1,g_2}_{\theta_{1l},\theta_{2l},\psi_l}\sharp\nu;c)\right]^{\frac{1}{2}},
\end{aligned}$$

which completes the proof.

B Additional Materials

HHRT with more than two marginals. We now extend the definition of HHRT to $K>2$ mariginals.

Definition 4 (Hierarchical Hybrid Radon Transform)**.** *Given $K\geq 2$, given defining functions $\{g_k:\mathbb{R}^{d_k}\times\Omega_i\to\mathbb{R}\}_{i=k}^K$, the Hierarchical Hybrid Radon Transform $\mathcal{HHR}:\mathbb{L}_1(\mathbb{R}^{d_1}\times\ldots\times\mathbb{R}^{d_K})\to\mathbb{L}_1\left(\mathbb{R}\times\Omega_1\ldots\times\Omega_K\times\mathbb{S}^{K-1}\right)$ is defined as:*

$$\begin{aligned}
&(\mathcal{HHR}f)(t,\theta_1,\ldots,\theta_K,\psi) \\
&\qquad=\int_{\mathbb{R}^{d_1}\times\ldots\times\mathbb{R}^{d_K}} f(x_1,\ldots,x_K)\delta\left(t-\sum_{k=1}^{K}\psi_k g_k(x_k,\theta_k)\right)dx_1\ldots dx_K. \qquad (12)
\end{aligned}$$

H2SW with more than two marginals. From the new definition of HRRT on $K>2$ mariginals, we now can define H2SW between joint distributions with K mariginals.

Definition 5. *For $p \geq 1, K \geq 2$, defining functions $g_1, \ldots, g_K$, the hierarchical hybrid sliced Wasserstein-p (H2SW) distance between two distributions $\mu \in \mathcal{P}(\mathcal{X}_1 \times \ldots \times \mathcal{X}_K)$ and $\nu \in \mathcal{P}(\mathcal{Y}_1 \times \ldots \times \mathcal{Y}_K)$ with an one-dimensional ground metric $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is defined as:*

$$H2SW_p^p(\mu, \nu; c, g_1, \ldots, g_K) = \mathbb{E}_{(\theta_1,\ldots,\theta_K,\psi)\sim\mathcal{U}(\Omega_1\times\ldots\times\Omega_K\times\mathbb{S}^{K-1})}[W_p^p(\mathcal{HHR}^{g_1,\ldots,g_K}_{\theta_1,\ldots,\theta_K,\psi}\sharp\mu, \mathcal{HHR}^{g_1,\ldots,g_K}_{\theta_1,\ldots,\theta_K,\psi}\sharp\nu; c)], \tag{13}$$

where $\mathcal{HHR}^{g_1,\ldots,g_K}_{\theta_1,\ldots,\theta_K,\psi}\sharp\mu$ and $\mathcal{HHR}^{g_1,\ldots,g_K}_{\theta_1,\ldots,\theta_K,\psi}\sharp\nu$ are the one-dimensional push-forward distributions created by applying HHRT.

Lorentz Model and Busemann function. The Lorentz model $\mathbb{L}^d \in \mathbb{R}^{d+1}$ of a d-dimensional hyperbolic space is [7]:

$$\mathbb{L}^d = \left\{(x_1, \ldots, x_d) \in \mathbb{R}^{d+1}, -x_0y_0 + \sum_{i=1}^{d} x_iy_i = -1, x_0 > 0\right\}.$$

Given a direction $\theta \in T_{x_0}\mathbb{L}^d \cap \mathbb{S}^d$, $x \in \mathbb{L}^d$, the Busemann function is:

$$B(x, \theta) = \log(-\langle x, x_0 + \theta\rangle).$$

Busemann function on product Hadamard manifolds. For distributions supports on the product of $K \geq 2$ Hadamard manifolds with the corresponding Busemann functions $B_1, \ldots, B_K$, we have a Buseman function of the product manifolds is:

$$B(x_1, \ldots, x_K, \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \lambda_k B_k(x_k, \theta_k),$$

for $(\lambda_1, \ldots, \lambda_K) \in \mathbb{S}^{K-1}$. The Cartan-Hyperbolic Sliced-Wasserstein distance use a fixed value of $(\lambda_1, \ldots, \lambda_K)$ e.g., $(\lambda_1, \ldots, \lambda_K) = (1/\sqrt{K}, \ldots, 1/\sqrt{K})$ (see [1]). In our proposed H2SW, we treat $(\lambda_1, \ldots, \lambda_K)$ as a random variable follows $\mathcal{U}(\mathbb{S}^{K-1})$ and the value of H2SW is defined as the mean of such random variable.

C Related Works

HHRT and Generalized Radon Transform. HHRT can be also seen as a special case of GRT [3] with the defining function $g(x, \theta) = \psi_1 g_1(x_1, \theta_1) + \psi_2 g_2(y, \theta_2)$ with $x = (x_1, x_2)$ and $\theta = (\theta_1, \theta_2, \psi)$ $(\Omega = \Omega_1 \times \Omega_2 \times \mathbb{S})$. However, without approaching via the hierarchical construction, the injectivity of the transform might be a challenge to obtain.

HHRT and Hierarchical Radon Transform. Hierarchical Radon Transform (HRT) [44] is the composition of Partial Radon Transform and Overparameterized Radon Transform, which is designed specifically for reducing projection complexity when using Monte Carlo estimation. Moreover, HRT is introduced with linear projection and does not focus on the problem of comparing heterogeneous

[1] https://github.com/clbonet/Sliced-Wasserstein_Distances_and_Flows_on_Cartan-Hadamard_Manifolds/blob/0eb05450e7f9f27586d0ddb1ce6e58f07eb75786/Experiments/xp_otdd/OTDD_SW.ipynb

Table 4: Summary of joint Wasserstein distances (multiplied by 10^2) across time steps.

Distances	Step 100 ($W_{c_1,c_2} \downarrow$)	Step 600 ($W_{c_1,c_2} \downarrow$)	Step 2000 ($W_{c_1,c_2} \downarrow$)	Step 3000 ($W_{c_1,c_2} \downarrow$)	Step 4000 ($W_{c_1,c_2} \downarrow$)	Step 5000 ($W_{c_1,c_2} \downarrow$)
SW L=10	1857.047±3.413	911.945±5.011	44.816±0.442	21.241±1.048	7.512±0.527	2.968±0.428
GSW L=10	1897.592±2.594	1033.613±6.449	73.957±1.846	30.212±1.241	9.375±0.088	4.756±0.126
H2SW L=10	**1847.303±2.242**	**906.201±4.694**	**43.562±1.928**	**19.173±1.172**	**6.094±0.172**	**1.931±0.194**
SW L=100	1851.449±0.137	898.554±1.279	39.203±1.968	17.198±1.119	4.832±0.472	1.47±0.411
GSW L=100	1892.182±0.64	1017.264±1.943	59.949±0.101	23.736±0.669	7.9±0.048	3.56±0.299
H2SW L=100	1843.801±1.275	888.824±2.197	39.217±0.913	17.169±0.432	5.33±0.228	1.404±0.139

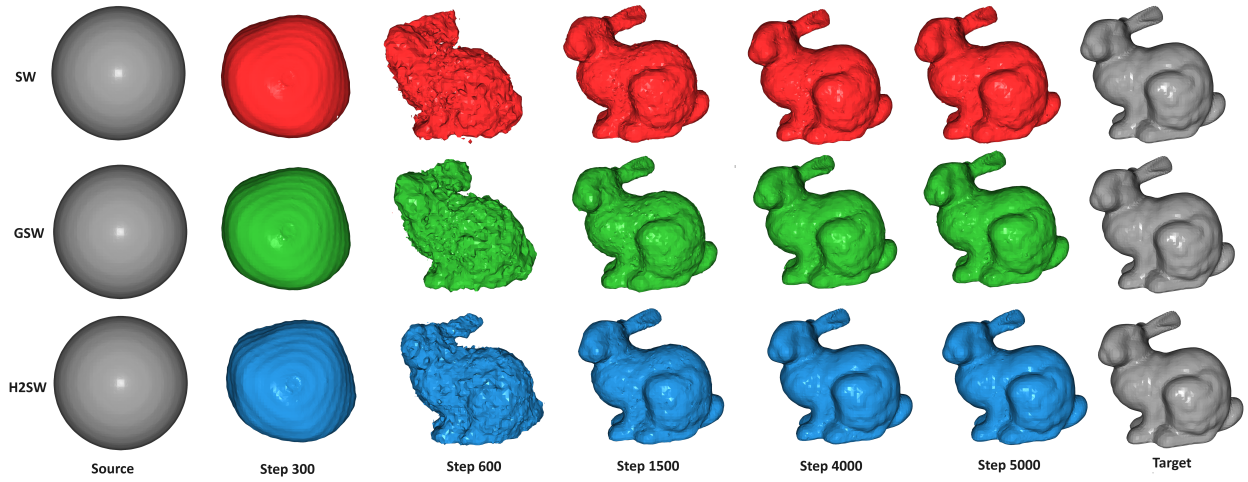

Figure 5: Visualization of deformation from the sphere mesh to the Stanford Bunny mesh with $L = 100$.

joint distributions. In contrast to HRT, the proposed HHRT is the composition of multiple partial Generalized Radon Transform and Partial Random Transform, which is suitable for comparing heterogeneous joint distributions.

HHRT and convolution slicers. Convolution slicers [41] are introduced to project an image into a scalar. It can be viewed as a Hierarchial Partial Radon Transform i.e., small parts of the image are transformed first, then be aggregated later. Although convolution slicers can separate global and local information as HHRT, they focus on the domain of images only and have not been proven to be injective. Again, HHRT is designed to compare heterogeneous joint distributions and is proven to be injective in Proposition 2. As a result, H2SW is a valid metric while convolution sliced Wasserstein [41] is only a pseudo metric. Moreover, H2SW can also use convolution slicers when having marginal domains as images.

D Additional Experiments

3D Mesh Deformation. As mentioned in the main text, we present the deformation visualization to the Stanford Bunny mesh with $L = 100$ in Figure 5, and the deformation visualization to the Armadillo mesh with $L = 10$ and $L = 100$ in Figure 6- 7 in turn. The quantitative result for the Armadillo mesh is given in Table 4. From these results, we see that the proposed H2SW gives the best flow deformation flow. The performance gap is especially larger when $L = 10$ i.e., having a small number of projections.

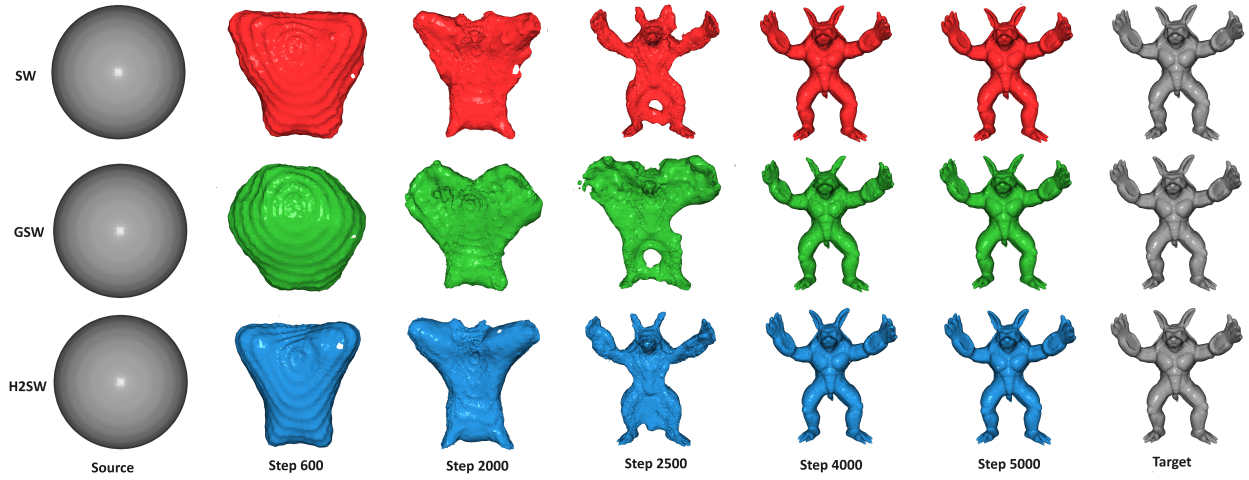

Figure 6: Visualization of deformation from the sphere mesh to the Armadillo mesh with $L = 10$.

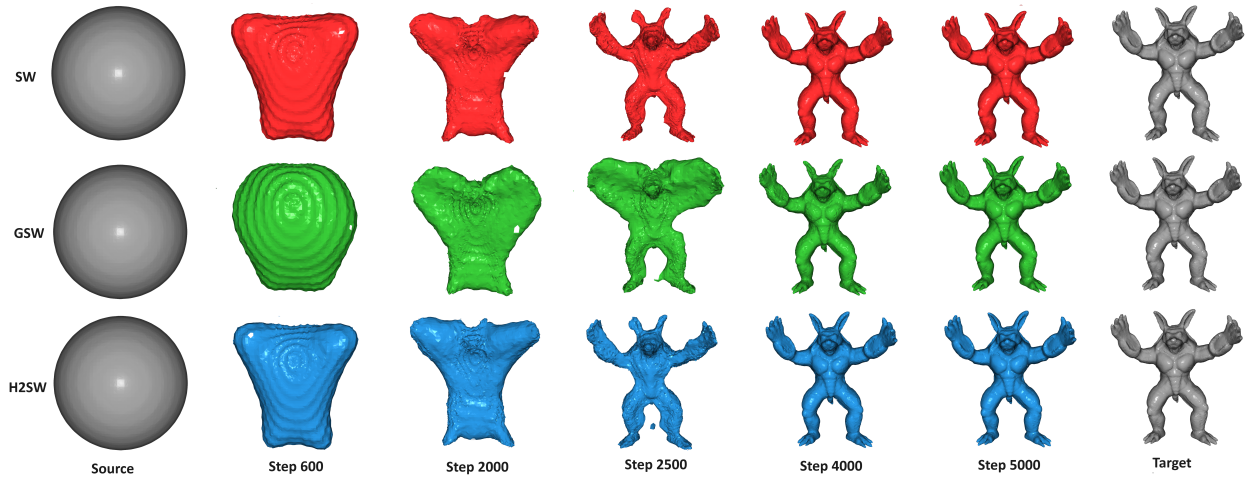

Figure 7: Visualization of deformation from the sphere mesh to the Armadillo mesh with $L = 100$.

Deep 3D mesh autoencoder. We first report the neural network architectures that we use in the experiments.

- The encoder: Conv1d(6,64,1) → BatchNorm1d → LeakyReLU(0.2) → Conv1d(64, 128, 1) → BatchNorm1d → LeakyReLU(0.2) → Conv1d(128, 256, 1) → BatchNorm1d → LeakyReLU(0.2) → Conv1d(256, 512, 1) → BatchNorm1d → LeakyReLU(0.2) → Conv1d(512, 1024, 1) → BatchNorm1d → LeakyReLU(0.2) → Max-Pooling → Linear(1024, 1024).

- The decoder: Linear(1024, 1024) → BatchNorm1d → LeakyReLU(0.2) → Linear(1024, 2048) → BatchNorm1d → LeakyReLU(0.2) → Linear(2048, 4096) → BatchNorm1d → LeakyReLU(0.2) → Linear(2048, 2048*6). The output of the decoder is the concatenation of the location and normal vector. We normalize the normal vector to the unit-sphere.

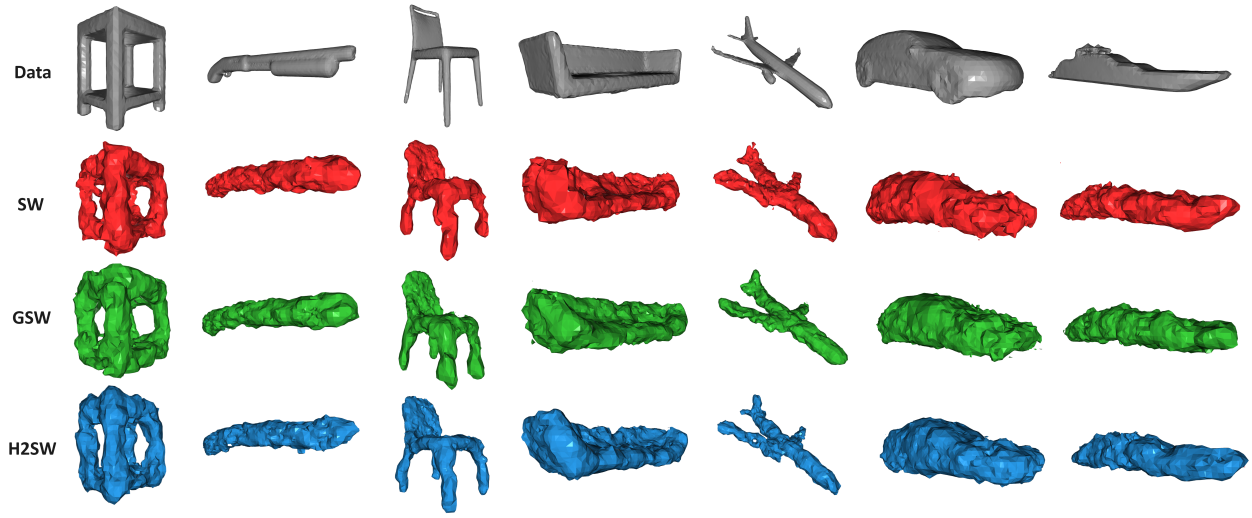

Figure 8: Visualization of some randomly selected reconstruction meshes from autoencoders trained by SW, GSW, and H2SW in turn with the number of projections $L = 100$ at epoch 500.

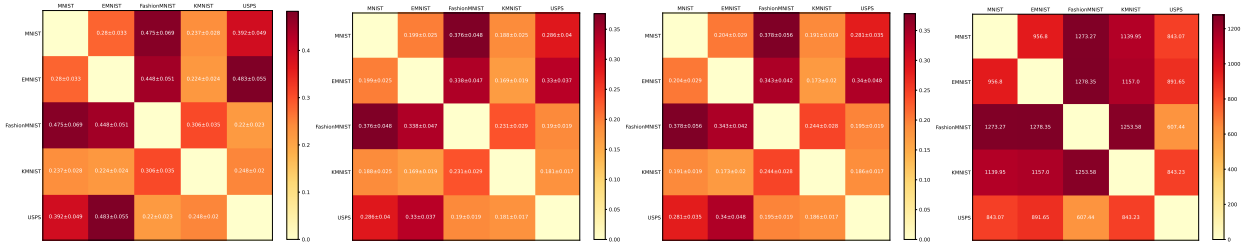

Figure 9: Cost matrices between datasets from SW, CHSW, and H2SW with $L = 100$.

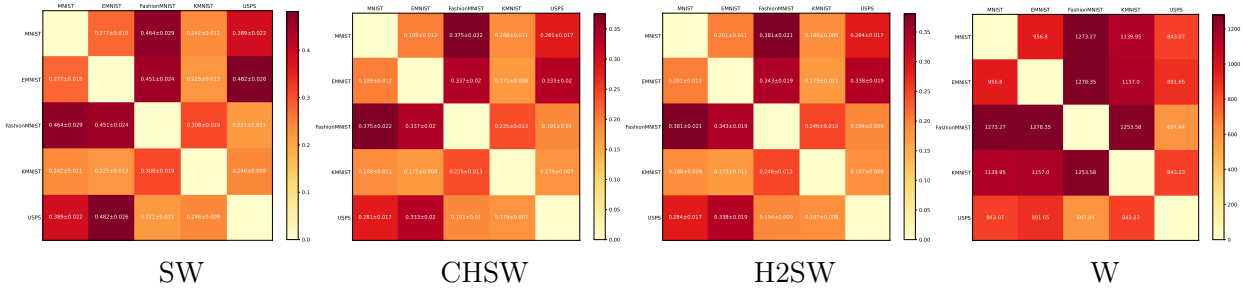

Figure 10: Cost matrices between datasets from SW, CHSW, and H2SW with $L = 500$.

As mentioned in the main text, we report the reconstruction of randomly selected meshes for $L = 100$ at epoch 500 in Figure 8. We see that the reconstructed meshes at epoch 500 are visually worse than the reconstructed meshes at epoch 2000. Therefore, the joint Wasserstein distances in Table 2 are consistent with the qualitative results.

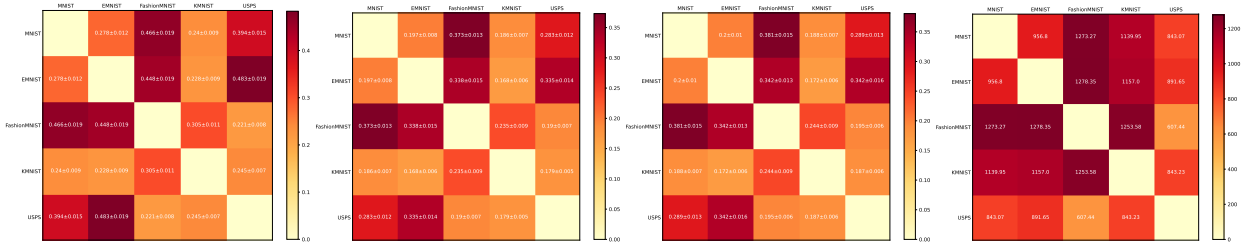

Figure 11: Cost matrices between datasets from SW, CHSW, and H2SW with $L = 1000$.

Dataset Comparison. We follow the same procedure in Section 6.2 in [9]. We refer the reader to the reference for a detailed description. Here, we show the cross-dataset cost matrices with the number of projections $L = 100$ in Figure 9, $L = 500$ in Figure 10, and $L = 1000$ Figure 11.

E Computational Infrastructure

For the non-deep-learning experiments, we use a HP Omen 25L desktop for conducting experiments. For 3D mesh autoencoder experiments, we use a single NVIDIA A100 GPU.

References

[1] D. Alvarez-Melis and N. Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020. (Cited on pages 2, 4, and 12.)

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. (Cited on page 1.)

[3] G. Beylkin. The inversion problem and applications of the generalized Radon transform. *Communications on pure and applied mathematics*, 37(5):579–599, 1984. (Cited on pages 2, 4, 5, and 19.)

[4] B. Bhushan Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018. (Cited on page 2.)

[5] M. T. Boedihardjo. Sharp bounds for the max-sliced wasserstein distance. *arXiv preprint arXiv:2403.00666*, 2024. (Cited on page 7.)

[6] C. Bonet, P. Berg, N. Courty, F. Septier, L. Drumetz, and M.-T. Pham. Spherical sliced-Wasserstein. *International Conference on Learning Representations*, 2023. (Cited on pages 2 and 5.)

[7] C. Bonet, L. Chapel, L. Drumetz, and N. Courty. Hyperbolic sliced-Wasserstein via geodesic and horospherical projections. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 334–370. PMLR, 2023. (Cited on pages 2, 8, and 19.)

[8] C. Bonet, N. Courty, F. Septier, and L. Drumetz. Efficient gradient flows in sliced-Wasserstein space. *Transactions on Machine Learning Research*, 2022. (Cited on page 2.)

[9] C. Bonet, L. Drumetz, and N. Courty. Sliced-Wasserstein distances and flows on Cartan-Hadamard manifolds. *arXiv preprint arXiv:2403.06560*, 2024. (Cited on pages 2, 8, 11, and 23.)

[10] C. Bonet, B. Malézieux, A. Rakotomamonjy, L. Drumetz, T. Moreau, M. Kowalski, and N. Courty. Sliced-Wasserstein on symmetric positive definite matrices for m/eeg signals. In *International Conference on Machine Learning*, pages 2777–2805. PMLR, 2023. (Cited on pages 2 and 8.)

[11] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 1(51):22–45, 2015. (Cited on pages 1 and 3.)

[12] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. (Cited on page 10.)

[13] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018. (Cited on page 11.)

[14] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017. (Cited on page 11.)

[15] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3730–3739, 2017. (Cited on pages 2 and 4.)

[16] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. (Cited on page 9.)

[17] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013. (Cited on page 1.)

[18] I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A. G. Schwing. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019. (Cited on page 8.)

[19] I. Deshpande, Z. Zhang, and A. G. Schwing. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3483–3491, 2018. (Cited on page 2.)

[20] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015. (Cited on page 1.)

[21] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738, 2015. (Cited on page 17.)

[22] R. C. Garrett, T. Harris, B. Li, and Z. Wang. Validating climate models with spherical convolutional Wasserstein distance. *arXiv preprint arXiv:2401.14657*, 2024. (Cited on page 8.)

[23] E. Heitz, K. Vanhoey, T. Chambon, and L. Belcour. A sliced Wasserstein loss for neural texture synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9412–9420, 2021. (Cited on page 2.)

[24] S. Helgason. The Radon transform on r n. In *Integral Geometry and Radon Transforms*, pages 1–62. Springer, 2011. (Cited on pages 2, 3, and 5.)

[25] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994. (Cited on page 11.)

[26] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, page 0, 2006. (Cited on page 5.)

[27] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 261–272, 2019. (Cited on pages 2 and 4.)

[28] S. Kolouri, G. K. Rohde, and H. Hoffmann. Sliced Wasserstein distance for learning Gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3427–3436, 2018. (Cited on page 2.)

[29] P. Kuchment. Generalized transforms of Radon type and their applications. In *Proceedings of Symposia in Applied Mathematics*, volume 63, page 67, 2006. (Cited on pages 2, 4, and 5.)

[30] T. Le, K. Nguyen, S. Sun, K. Han, N. Ho, and X. Xie. Diffeomorphic mesh deformation via efficient optimal transport for cortical surface reconstruction. *International Conference on Learning Representations*, 2024. (Cited on page 2.)

[31] T. Le, K. Nguyen, S. Sun, N. Ho, and X. Xie. Integrating efficient optimal transport and functional maps for unsupervised shape correspondence learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. (Cited on page 2.)

[32] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. (Cited on page 11.)

[33] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht. Sliced Wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019. (Cited on page 2.)

[34] Z.-P. Liang and D. C. Munson. Partial Radon transforms. *IEEE transactions on image processing*, 6(10):1467–1469, 1997. (Cited on page 5.)

[35] A. Liutkus, U. Simsekli, S. Majewski, A. Durmus, and F.-R. Stöter. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pages 4104–4113. PMLR, 2019. (Cited on page 2.)

[36] M. Luong, K. Nguyen, N. Ho, R. Haf, D. Phung, and L. Qu. Revisiting deep audio-text retrieval through the lens of transportation. In *The Twelfth International Conference on Learning Representations*, 2024. (Cited on page 1.)

[37] T. Manole, S. Balakrishnan, and L. Wasserman. Minimax confidence intervals for the sliced Wasserstein distance. *Electronic Journal of Statistics*, 16(1):2252–2345, 2022. (Cited on page 7.)

[38] K. Nadjahi, V. De Bortoli, A. Durmus, R. Badeau, and U. Şimşekli. Approximate Bayesian computation with the sliced-Wasserstein distance. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5470–5474. IEEE, 2020. (Cited on page 2.)

[39] K. Nadjahi, A. Durmus, L. Chizat, S. Kolouri, S. Shahrampour, and U. Simsekli. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020. (Cited on page 7.)

[40] K. Nguyen, N. Bariletto, and N. Ho. Quasi-monte carlo for 3d sliced Wasserstein. In *The Twelfth International Conference on Learning Representations*, 2024. (Cited on page 7.)

[41] K. Nguyen and N. Ho. Revisiting sliced Wasserstein on images: From vectorization to convolution. *Advances in Neural Information Processing Systems*, 2022. (Cited on pages 8 and 20.)

[42] K. Nguyen and N. Ho. Energy-based sliced Wasserstein distance. *Advances in Neural Information Processing Systems*, 2023. (Cited on page 8.)

[43] K. Nguyen, N. Ho, T. Pham, and H. Bui. Distributional sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021. (Cited on page 8.)

[44] K. Nguyen, T. Ren, H. Nguyen, L. Rout, T. Nguyen, and N. Ho. Hierarchical sliced Wasserstein distance. *International Conference on Learning Representations*, 2023. (Cited on page 19.)

[45] S. Nietert, R. Sadhu, Z. Goldfeld, and K. Kato. Statistical, robustness, and computational guarantees for sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 2022. (Cited on page 7.)

[46] O. Pele and M. Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, September 2009. (Cited on page 1.)

[47] S. Peng, C. Jiang, Y. Liao, M. Niemeyer, M. Pollefeys, and A. Geiger. Shape as points: A differentiable poisson solver. *Advances in Neural Information Processing Systems*, 34:13032–13044, 2021. (Cited on pages 5 and 10.)

[48] G. Peyré and M. Cuturi. Computational optimal transport, 2020. (Cited on pages 1, 3, and 14.)

[49] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. (Cited on pages 10 and 11.)

[50] M. Quellmalz, R. Beinert, and G. Steidl. Sliced optimal transport on the sphere. *arXiv preprint arXiv:2304.09092*, 2023. (Cited on pages 2 and 5.)

[51] M. Quellmalz, L. Buecher, and G. Steidl. Parallelly sliced optimal transport on spheres and on the rotation group. *arXiv preprint arXiv:2401.16896*, 2024. (Cited on pages 2 and 5.)

[52] F. Rouviere. Nonlinear Radon and Fourier transforms, 2015. (Cited on pages 2 and 4.)

[53] R. M. Rustamov and S. Majumdar. Intrinsic sliced Wasserstein distances for comparing collections of probability distributions on manifolds and graphs. In *International Conference on Machine Learning*, pages 29388–29415. PMLR, 2023. (Cited on page 2.)

[54] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018. (Cited on page 1.)

[55] H. Tran, Y. Bai, A. Kothapalli, A. Shahbazi, X. Liu, R. D. Martin, and S. Kolouri. Stereographic spherical sliced Wasserstein distances. *arXiv preprint arXiv:2402.02345*, 2024. (Cited on pages 2 and 5.)

[56] C. Villani. *Optimal transport: Old and New*. Springer, 2008. (Cited on page 1.)

[57] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. (Cited on page 3.)

[58] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. (Cited on page 11.)

[59] J. Xu, H. Zhou, C. Gan, Z. Zheng, and L. Li. Vocabulary learning via optimal transport for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7361–7373, 2021. (Cited on page 1.)

[60] M. Yi and S. Liu. Sliced Wasserstein variational inference. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2021. (Cited on page 2.)

[61] Q.-Y. Zhou, J. Park, and V. Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. (Cited on page 9.)