SLICING WASSERSTEIN OVER WASSERSTEIN VIA FUNCTIONAL OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Wasserstein distances define a metric between probability measures on arbitrary metric spaces, including *meta-measures* (measures over measures). The resulting *Wasserstein over Wasserstein* (WoW) distance is a powerful, but computationally costly tool for comparing datasets or distributions over images and shapes. Existing sliced WoW accelerations rely on parametric meta-measures or the existence of high-order moments, leading to numerical instability. As an alternative, we propose to leverage the isometry between the 1d Wasserstein space and the quantile functions in the function space $L_2([0,1])$. For this purpose, we introduce a general sliced Wasserstein framework for arbitrary Banach spaces. Due to the 1d Wasserstein isometry, this framework defines a sliced distance between 1d meta-measures via infinite-dimensional L_2-projections, parametrized by Gaussian processes. Combining this 1d construction with classical integration over the Euclidean unit sphere yields the *double-sliced Wasserstein* (DSW) metric for general meta-measures. We show that DSW minimization is equivalent to WoW minimization for discretized meta-measures, while avoiding unstable higher-order moments and computational savings. Numerical experiments on datasets, shapes, and images validate DSW as a scalable substitute for the WoW distance.

1 INTRODUCTION

Optimal transport (OT) enables geometrically meaningful Wasserstein distances between probability measures on arbitrary Polish spaces $\mathcal{X}$ (Villani, 2003), while remaining computationally tractable. This has had a profound impact on machine learning, where Wasserstein distances are used to train neural networks (Tong et al., 2024) and to compare data distributions (Yang et al., 2019). A key feature of OT is its applicability to non-Euclidean spaces, even allowing the definition of Wasserstein distances on Wasserstein spaces $\mathcal{P}_2(\mathcal{X})$ (Bonet et al., 2025b). This is particularly useful for comparing distributions over non-Euclidean objects. For example, Euclidean distances as ground metric between two images often yield poor results (Stanczuk et al., 2021), whereas Wasserstein distances are robust to small image perturbations (Beckmann et al., 2025). Similarly, comparing point clouds (Nguyen et al., 2021) is natural with OT but not even well-defined with Euclidean distances. While most OT applications focus either on comparing pairwise objects or distributions over Euclidean spaces, recent work leverages Wasserstein distances on Wasserstein spaces for non-Euclidean domains, such as image (Dukler et al., 2019) or point cloud (Haviv et al., 2025) distributions.

The underlying concept of multilevel OT has been introduced under various names, including hierarchical OT (Schmitzer & Schnörr, 2013; Lee et al., 2019), mixture Wasserstein (Chen et al., 2018; 2019; Delon & Desolneux, 2020), and Wasserstein over Wasserstein (WoW) (Bonet et al., 2025b). It has applications beyond images and shapes, including domain adaptation (Lee et al., 2019; El Hamri et al., 2022), single-cell analysis (Lin et al., 2023), point cloud registration (Steuernagel et al., 2023), Bayesian inference (Nguyen & Mueller, 2024), generative modelling (Atanackovic et al., 2025; Haviv et al., 2025), document analysis (Yurochkin et al., 2019), Gromov–Wasserstein approximations (Mémoli, 2011; Piening & Beinert, 2025a), and reinforcement learning (Ziesche & Rozo, 2023). Extending this framework with another Polish space $\mathcal{Y}$ for dataset labels yields the OT dataset distance (OTDD), defined on $\mathcal{P}_2(\mathcal{Y} \times \mathcal{P}_2(\mathcal{X}))$ (Alvarez-Melis & Fusi, 2020). However, all these approaches incur high computational cost due to repeated pairwise Wasserstein evaluations.

Due to the complexity of OT, sliced Wasserstein distances (Bonneel et al., 2015) provide efficient OT-based alternatives to standard Wasserstein distances. Initially developed for probability measures on Euclidean spaces, they have since been extended to the sphere (Bonet et al., 2023a; Quellmalz et al., 2023), manifolds (Bonet et al., 2025a), functions (Garrett et al., 2024), hyperbolic spaces (Bonet et al., 2023b), the rotation group (Quellmalz et al., 2024), and matrices (Bonet et al., 2023c). For WoW-type distances, sliced accelerations have been proposed for Gaussian mixtures (Nguyen & Mueller, 2024; Piening & Beinert, 2025b) and more generally for measures over measures (*meta-measures*) via the sliced OTDD (s-OTDD) (Nguyen et al., 2025). The s-OTDD employs a hierarchical slicing approach based on the method of moments. However, it is only well-defined for finite moments, and practical implementations are limited to the first few moments because of numerical instability – originally, the first five.

In this paper, we aim to circumvent the issues of the s-OTDD. Therefore, we build on theoretical ideas originally proposed in (Han, 2023) to develop a computable sliced Wasserstein metric on general Banach spaces. Employing the isometry between 1d probability measures in the Wasserstein space and quantile functions embedded in the space of square-integrable functions $L_2([0,1])$, we utilize this to define a sliced Wasserstein metric on the space of 1d meta-measures via L_2-projections. Due to the lack of a uniform distribution on the unit ball of infinite-dimensional function spaces, we parametrize our projection directions as Gaussian processes (Kanagawa et al., 2018). To extend this idea to multi-dimensional meta-measures, we combine this approach with a classical slicing approach, mapping these meta-measures to 1d meta-measures. Lastly, we prove that the minimization of our sliced distance results in WoW minimization. This leads to the following contributions:

- We generalize the sliced Wasserstein distance to arbitrary Banach spaces. Moreover, we show how two distinct parameterizations of the random projections may result in equivalent metrics. As a special case, this allows for a sliced distance between 1d meta-measures.
- Beyond 1d meta-measures, we extend our approach to the multivariate case by introducing the double-sliced Wasserstein (DSW) metric between meta-measures. Illustrating the usefulness of our DSW metric as a WoW replacement, we prove a form of topological metric equivalence between the two for discretized meta-measures.
- Lastly, we present various numerical experiments showcasing the advantages of our approach, allowing for meaningful distribution comparisons for datasets, shapes, and images.

2 Wasserstein Distances

The so-called Wasserstein distance or Kantorovich–Rubinstein metric is an optimal transport-based similarity gauge between probabilities on a common Polish space. To this end, let $\mathcal{X}$ be a Polish space, let $\mathcal{P}(\mathcal{X})$ be the space of Borel probability measures on $\mathcal{X}$ with respect to the Borel σ-algebra induced by the underlying metric, and let $\mathcal{P}_p(\mathcal{X})$, $p \in [1,\infty)$, be the subset of measures with finite pth moment. For $\mu \in \mathcal{P}(\mathcal{X})$ and a second Polish space $\mathcal{Y}$, the *push-forward* by a measurable map $T\colon \mathcal{X} \to \mathcal{Y}$ is defined by $T_\sharp\, \mu := \mu \circ T^{-1} \in \mathcal{P}(\mathcal{Y})$. The set of *transport plans* between $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$ is given by

$$\Gamma(\mu,\nu) := \big\{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \bigm| \pi_{1,\sharp}\, \gamma = \mu, \pi_{2,\sharp}\, \gamma = \nu\big\},$$

where π_i denotes the canonical projection onto the ith component. For a complete, separable metric space $(\mathcal{X}, d)$, the *(2-)Wasserstein distance*

$$\mathrm{W}(\mu,\nu;\mathcal{X}) := \inf_{\gamma \in \Gamma(\mu,\nu)} \Big(\int_{\mathcal{X}\times\mathcal{X}} d^2(x_1,x_2)\ \mathrm{d}\gamma(x_1,x_2)\Big)^{\frac{1}{2}}$$

defines a metric on $\mathcal{P}_2(\mathcal{X})$. More precisely, $(\mathcal{P}_2(\mathcal{X}), \mathrm{W})$ is again a complete separable metric space, allowing the construction of the so-called *Wasserstein over Wasserstein* (WoW) *distance* $\mathbf{W}(\cdot,\cdot;\mathcal{X}) := \mathrm{W}(\cdot,\cdot;\mathcal{P}_2(\mathcal{X}))$, which is studied in (Bonet et al., 2025b).

In difference to other similarity gauges like the Kullback–Leibler divergence or the total variation, Wasserstein distances leverage the underlying geometry, allowing for meaningful comparisons between empirical measures. Although the Wasserstein distance relies on a linear program, the actual calculation is computationally costly. For two empirical measures supported at n points in $\mathbb{R}^d$ equipped with the Euclidean metric, the exact computation has complexity $\mathcal{O}(n^3 \log n)$. The approximate computation based on entropic regularization still has complexity $\mathcal{O}(n^2 \log n)$, see (Peyré & Cuturi, 2019). Notably, this computational burden becomes even more involved for non-Euclidean metric spaces, where the computation of the underlying distance itself is challenging. For instance, the computation of the WoW distance relies on the pointwise evaluations of Wasserstein distances. If the empirical meta-measures in $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ are supported on N empirical measures on $\mathcal{P}_2(\mathbb{R}^d)$, each with n support points, then the approximate calculation of the required distance matrix already has complexity $\mathcal{O}(N^2 n^2 \log n)$.

From a computational point of view, the Wasserstein distance on $(\mathbb{R}, |\cdot - \cdot|)$ is a notable exception since this may be evaluated analytically. To this end, for $\mu \in \mathcal{P}(\mathbb{R})$, its *quantile function* $Q_\mu \colon (0,1) \to \mathbb{R}$ is given by $Q_\mu(s) \coloneqq \inf \big\{ x \in \mathbb{R} \;\big|\; \mu((-\infty, x]) \geq s \big\}$. The Wasserstein distance now becomes

$$\mathrm{W}(\mu, \nu; \mathbb{R}) = \left(\int_0^1 |Q_\mu(s) - Q_\nu(s)|^2 \, \mathrm{d}s \right)^{\frac{1}{2}}, \quad \mu, \nu \in \mathcal{P}_2(\mathbb{R}),$$

meaning that the *quantile mapping*

$$q \colon \mathcal{P}_2(\mathbb{R}) \to L_2([0,1]), \quad \mu \mapsto Q_\mu \tag{1}$$

is an isometric embedding into the space of square-integrable functions, see (Villani, 2003). For empirical measures, the quantile functions are piecewise constant and can be efficiently computed by sorting the support points.

3 Sliced Wasserstein Distances

At their core, all sliced Wasserstein distances exploit easy-to-compute, 1d optimal transports to define efficient alternatives to the standard Wasserstein distance. Originally, the sliced Wasserstein distance has been studied for measures in $\mathcal{P}_2(\mathbb{R}^d)$ and is based on the *slicing operator*

$$\pi_\theta \colon \mathbb{R}^d \to \mathbb{R}, \; x \mapsto \langle \theta, x \rangle, \qquad \theta \in \mathbb{S}^{d-1} \coloneqq \{ x \in \mathbb{R}^d \mid \|x\| = 1 \}, \tag{2}$$

with respect to the Euclidean inner product and norm. The *sliced (2-)Wasserstein distance* reads as

$$\mathrm{SW}(\mu, \nu) \coloneqq \left(\int_{\mathbb{S}^{d-1}} \mathrm{W}^2(\pi_{\theta,\sharp}\, \mu, \pi_{\theta,\sharp}\, \nu; \mathbb{R}) \, \mathrm{d}\mathbb{S}^{d-1}(\theta) \right)^{\frac{1}{2}}, \tag{3}$$

where we integrate with respect to the uniform probability on $\mathbb{S}^{d-1}$. Similar to the Wasserstein distance, SW metricizes the weak convergence (Bonnotte, 2013; Nadjahi et al., 2020). The spherical integral is usually approximated by Monte Carlo schemes (Bonneel et al., 2015; Nguyen et al., 2024; Hertrich et al., 2025) or Gaussian approximation (Nadjahi et al., 2021).

3.1 Slicing Infinite Dimensional Banach Spaces

As preliminary step towards an implementable sliced WoW distance, we consider the slicing on an infinite dimensional, separable Banach space U with dual U^*. Relying on the dual pairing, we generalize the slicing operator (2) by

$$\pi_v \colon U \to \mathbb{R}, \; u \mapsto \langle v, u \rangle, \qquad v \in U^*.$$

The crucial point in defining a sliced distance on $\mathcal{P}_2(U)$ is that there exists no uniform probability measure on the infinite dimensional sphere. As remedy, we choose an arbitrary $\xi \in \mathcal{P}_2(U^*)$ and define the ξ-based *sliced Wasserstein distance* as

$$\mathrm{SW}(\mu, \nu; \xi) \coloneqq \left(\int_{U^*} \mathrm{W}^2(\pi_{v,\sharp}\, \mu, \pi_{v,\sharp}\, \nu; \mathbb{R}) \, \mathrm{d}\xi(v) \right)^{\frac{1}{2}}, \quad \mu, \nu \in \mathcal{P}(U). \tag{4}$$

This approach extends the slicing on Hilbert spaces (Han, 2023). However, unlike (Han, 2023), we do not construct specific measures on the sphere. This allows the use of easy-to-sample slicing measures. If the support of ξ covers all directions in U^*, the ξ-based SW distance becomes a metric. The proof is given in Appendix A.

Theorem 3.1. *For $\xi \in \mathcal{P}_2(U^*)$, the ξ-based SW distance defines a well-defined pseudo-metric. If $\operatorname{supp}\xi \cap \operatorname{span} v \notin \big\{\emptyset, \{0\}\big\}$ for all $v \in U^*$, then (4) defines a metric on $\mathcal{P}_2(U)$.*

If the slicing measure ξ has full support, then the assumption in Theorem 3.1 is fulfilled, and the ξ-based SW distance is a metric. Two measures $\xi_1, \xi_2 \in \mathcal{P}_2(U^*)$ are equivalent if they are mutually absolutely continuous. If their Radon–Nikodým derivatives $\mathrm{d}\xi_1/\,\mathrm{d}\xi_2$ and $\mathrm{d}\xi_2/\,\mathrm{d}\xi_1$ are bounded, then the resulting SW distances are metrically equivalent. The proof is again given in Appendix A.

Proposition 3.2. *Let $\xi_1, \xi_2 \in \mathcal{P}_2(U^*)$ be equivalent. If $\mathrm{d}\xi_1/\,\mathrm{d}\xi_2$ and $\mathrm{d}\xi_2/\,\mathrm{d}\xi_1$ are bounded, then we find $c_1, c_2 > 0$ such that*

$$c_1 \,\mathrm{SW}(\mu,\nu;\, \xi_1) \le \mathrm{SW}(\mu,\nu;\, \xi_2) \le c_2 \,\mathrm{SW}(\mu,\nu;\, \xi_1) \quad \forall \mu,\nu \in \mathcal{P}_2(U).$$

In the finite-dimensional Euclidean setting, special cases of the ξ-based SW distance correspond, for instance, to so-called energy measures on the sphere (Nguyen & Ho, 2023) and the standard Gaussian (Nadjahi et al., 2021). Relying on the latter, we obtain a strong equivalence to original SW (3) if ξ is equivalent to the standard Gaussian. The short proof is given in Appendix A.

Proposition 3.3. *Let $\xi \in \mathcal{P}_2(\mathbb{R}^d)$ be equivalent to $\eta \sim \mathcal{N}(0, \mathbf{I}_d)$. If $\mathrm{d}\xi/\,\mathrm{d}\eta$ and $\mathrm{d}\eta/\,\mathrm{d}\xi$ are bounded, then we find $c_1, c_2 > 0$ such that*

$$c_1 \,\mathrm{SW}(\mu,\nu;\xi) \le \mathrm{SW}(\mu,\nu) \le c_2 \,\mathrm{SW}(\mu,\nu;\xi) \quad \forall \mu,\nu \in \mathcal{P}_2(\mathbb{R}^d).$$

Assuming that samples of ξ are available, the ξ-based SW distance on every separable Banach space can again be computed using the Monte Carlo scheme. If ξ has a finite fourth moment, i.e. $\xi \in \mathcal{P}_4(U^*)$, the Monte Carlo scheme converges. The details are given in Proposition A.6.

3.2 Slicing the 1D Wasserstein Space

Exploiting the generalized SW distance in (4), we introduce a first slicing of the Wasserstein space $(\mathcal{P}_2(\mathbb{R}), \mathrm{W})$, which later builds the foundation of our sliced WoW distance on $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$. Recall that the quantile mapping (1) is an isometric embedding and thus measurable. Therefore, we can push every meta-measure $\boldsymbol{\mu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}))$ to $q_\sharp\boldsymbol{\mu} \in \mathcal{P}_2(L_2([0,1]))$. In this manner, the WoW distance between $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}))$ becomes $\mathbf{W}(\boldsymbol{\mu},\boldsymbol{\nu};\mathbb{R}) = \mathrm{W}(q_\sharp\boldsymbol{\mu}, q_\sharp\boldsymbol{\nu}; \mathcal{P}_2(L_2([0,1]))$. Fixing $\xi \in \mathcal{P}_2(L_2([0,1]))$, we introduce the *sliced quantile WoW* (SQW) *distance:*

$$\mathbf{SQW}(\boldsymbol{\mu},\boldsymbol{\nu};\xi) \coloneqq \mathrm{SW}(q_\sharp\boldsymbol{\mu}, q_\sharp\boldsymbol{\nu};\, \xi), \qquad \boldsymbol{\mu},\boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R})). \tag{5}$$

If ξ is *positive*, i.e., ξ has full support, then the assumptions of Theorem 3.1 are satisfied.

Corollary 3.3.1. *Let $\xi \in \mathcal{P}_2(L_2([0,1]))$ be positive, then $\mathbf{SQW}$ is a metric on $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}))$.*

For the later implementation, we require an easy-to-sample slicing measure. To this end, we propose to use Gaussians, i.e., measures $\xi \in \mathcal{P}_2(L_2([0,1]))$ such that $\pi_{v,\sharp}\xi$ is Gaussian for all $v \in L_2([0,1])$, see (Bogachev, 1998). On $L_2([0,1])$, there exists a one-to-one correspondence between Gaussian measures and Gaussian processes (Rajput & Cambanis, 1972, Thm. 2). In our numerics, we restrict ourselves to the Gaussian process G that is related to the covariance kernel

$$k_\sigma \colon [0,1] \times [0,1] \to \mathbb{R}, \quad (t,s) \mapsto \exp(-|t-s|^2/2\sigma^2). \tag{6}$$

This means that we consider the function-valued random variable G with

$$\big(G(t_1), \dots, G(t_n)\big) \sim \mathcal{N}\big(0, (k(t_i,t_j))_{i,j=1,\dots,n}\big) \qquad \forall t_1,\dots,t_n \in [0,1].$$

Since the kernel is smooth, the sample paths (realizations) of G are smooth too, i.e., $G \in \mathcal{C}^\infty([0,1])$ almost surely, see (Costa et al., 2023, Cor. 1). Since k_σ is *universal* (Steinwart, 2001), the corresponding Gaussian measure has full support (Van Der Vaart et al., 2008).

4 Double-Slicing the Wasserstein Space

The slicing schemes in § 3.1 and § 3.2 cannot directly be generalized to the multidimensional Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W})$ due to the lack of the Banach space structure and since we do not have an adequate generalization of the quantile mapping. Instead of slicing the Wasserstein space directly, in a first step, we therefore propose to slice the underlying domain using

$$\boldsymbol{\pi}_\theta \colon \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}), \quad \mu \mapsto \pi_{\theta,\sharp}\mu, \qquad \theta \in \mathbb{S}^{d-1}.$$

with π_θ from (2). Notice that $\boldsymbol{\pi}_\theta$ is continuous with respect to the Wasserstein distances and thus measurable. This allows us to define the *sliced WoW distance* via

$$\mathbf{SW}(\boldsymbol{\mu}, \boldsymbol{\nu}) \coloneqq \Big(\int_{\mathbb{S}^{d-1}} \mathbf{W}^2(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu}, \boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\nu}; \mathbb{R}) \, \mathrm{d}\mathbb{S}^{d-1}(\theta) \Big)^{\frac{1}{2}}, \qquad \boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)),$$

where we again integrate with respect to the uniform measure. Based on finite-dimensional slicing, $\mathbf{SW}$ essentially reduces meta-measures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ to a series of 1d meta-measures $\pi_{\theta,\sharp}\boldsymbol{\mu}$ and $\pi_{\theta,\sharp}\boldsymbol{\nu}$.

As the computation of the 1d WoW distance remains challenging, we resort to a hierarchical slicing approach, inspired by (Nguyen et al., 2025; Piening & Beinert, 2025b). Using quantile slicing (5) with slicing measure $\xi \in \mathcal{P}_2(L_2([0,1]))$, we introduce the *double-sliced WoW distance*

$$\mathbf{DSW}(\boldsymbol{\mu}, \boldsymbol{\nu}; \xi) \coloneqq \Big(\int_{\mathbb{S}^{d-1}} \mathbf{SQW}^2(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu}, \boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\nu}; \xi) \, \mathrm{d}\mathbb{S}^{d-1}(\theta) \Big)^{\frac{1}{2}}, \qquad \boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)).$$

By construction, the double-sliced WoW distance defines at least a pseudo-metric. If we restrict ourselves to empirical meta-measure, $\mathbf{DSW}$ even becomes a metric that is weakly equivalent to $\mathbf{W}$. To be more precise, we denote the subset of empirical measures by $\mathcal{P}_\mathrm{e}$ and the Dirac measure by $\delta_\bullet$. An *empirical meta-measure* $\boldsymbol{\mu} \in \mathcal{P}_\mathrm{e}(\mathcal{P}_\mathrm{e}(\mathbb{R}^d))$ has the form

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\mu_i} \quad \text{with} \quad \mu_i = \frac{1}{n_i} \sum_{k=1}^{n_i} \delta_{x_{i,k}} \quad \text{and} \quad x_{i,k} \in \mathbb{R}^d \tag{7}$$

for arbitrary N and n_i. For N and $n_i \equiv \tilde{n}$ fixed, we denote $\boldsymbol{\mu} \in \mathcal{P}_\mathrm{e}^N(\mathcal{P}_\mathrm{e}^{\tilde{n}}(\mathbb{R}^d))$ with $\mu_i \in \mathcal{P}_\mathrm{e}^{\tilde{n}}(\mathbb{R}^d)$.

Theorem 4.1. *For positive $\xi \in \mathcal{P}_2(L_2([0,1]))$, $\mathbf{DSW}$ defines a metric on $\mathcal{P}_\mathrm{e}(\mathcal{P}_\mathrm{e}(\mathbb{R}^d))$. Moreover, for $\boldsymbol{\mu}_n, \boldsymbol{\mu} \in \mathcal{P}_\mathrm{e}^N(\mathcal{P}_\mathrm{e}^{\tilde{n}}(\mathcal{X}))$ with compact $\mathcal{X} \subset \mathbb{R}^d$ and positive Gaussian ξ, it holds*

$$\mathbf{DSW}(\boldsymbol{\mu}_n, \boldsymbol{\mu}; \xi) \to 0 \iff \mathbf{SW}(\boldsymbol{\mu}_n, \boldsymbol{\mu}; \xi) \to 0 \iff \mathbf{W}(\boldsymbol{\mu}_n, \boldsymbol{\mu}; \mathbb{R}^d) \to 0 \quad \textit{as} \quad n \to \infty.$$

Numerically, $\mathbf{DSW}$ can be implemented combining several integration techniques. In the following, we consider the Gaussian ξ related to the kernel in (6) and empirical meta-measures $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_\mathrm{e}(\mathcal{P}_\mathrm{e}(\mathbb{R}^d))$ as in (7). To approximate the outer integral over $\mathbb{S}^{d-1}$ and the inner integral over ξ simultaneously, we employ the Monte Carlo method. For a sample path g of the corresponding Gaussian process and a direction $\theta \in \mathbb{S}^{d-1}$, we evaluate the integrand as follows: The first slicing gives $\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu} = (1/N)\sum_{i=1}^N \delta_{\pi_{\theta,\sharp}\mu_i}$, and the quantile mapping $q_\sharp \boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu} = (1/N)\sum_{i=1}^N \delta_{q(\pi_{\theta,\sharp}\mu_i)}$, where the piecewise constant quantile functions can be determined by sorting the support points. For the slicing on $L_2([0,1])$, we employ a quadrature to approximate the inner product. To this end, for knots $t_1, \dots, t_R \in [0,1]$ and weights $w_1, \dots, w_R$, we estimate

$$\widehat{\pi_{g,\sharp} q_\sharp \boldsymbol{\pi}_{\theta,\sharp}\mu} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\widehat{\langle q(\pi_{\theta,\sharp}\mu_i), g \rangle}} \quad \text{with} \quad \widehat{\langle q(\pi_{\theta,\sharp}\mu_i), g \rangle} = \sum_{r=1}^{R} w_r \, q(\pi_{\theta,\sharp}\mu_i)(t_r) \, g(t_r).$$

Note that the samples g of the process G satisfy $(G(t_1), \dots, G(t_R)) \sim \mathcal{N}(0, k(t_r, t_{r'}))_{r,r'=1}^R$ and can be easily generated. Finally, the double-sliced WoW distance is computed by

$$\widehat{\mathrm{DSW}}(\boldsymbol{\mu}, \boldsymbol{\nu}) \coloneqq \Big(\frac{1}{S} \sum_{s=1}^{S} \mathrm{W}^2 \big(\widehat{\pi_{g_s,\sharp} q_\sharp \boldsymbol{\pi}_{\theta_s,\sharp}\mu}, \widehat{\pi_{g_s,\sharp} q_\sharp \boldsymbol{\pi}_{\theta_s,\sharp}\nu}; \mathbb{R} \big) \Big)^{\frac{1}{2}}.$$

The remaining 1d Wasserstein distances can again be efficiently computed.

5 Numerical Experiments

In this section, we aim to showcase the numerical properties and benefits of our sliced distances. We start with the 1d case. Drawing a connection between meta-measures and the so-called Gromov–Wasserstein (GW) distance, we consider a shape classification experiment from (Piening & Beinert, 2025a). Next, we compare our multidimensional sliced distance to the s-OTDD (Nguyen et al., 2025). Finally, we present applications from the evaluation of point cloud distributions and perceptual image analysis. For all these experiments, we employ trapezoidal integration weights w_r. We refer to Appendix C for further experiments and details.

5.1 Shape Classification via Local Distance Distributions

First, we repeat a shape classification experiment from (Piening & Beinert, 2025a) based on parametrizing shapes as so-called metric measure (mm-)spaces. A mm-space is a tuple $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ consisting of a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$ and a measure $\mu \in \mathcal{P}_2(\mathcal{X})$. Modelling data as (finite) metric spaces allows for invariance to isometric transformations such as rotations, often desirable for shapes. Particularly, we may parametrize 2d shapes as point clouds with pairwise Euclidean distances and 3d shapes as triangular meshes with pairwise surface distances (Beier et al., 2022).

While Gromov–Wasserstein (GW) distances define a metric between mm–spaces (Mémoli, 2011), computation is costly and relies on inexact non-convex minimization. As a remedy, alternatives employ pseudo-metrics via *local distance distributions*. Namely, they map a finite, uniformly-weighted mm-space $\mathbb{X} = (x_1, \ldots, x_N)$ with $\mu \in \mathcal{P}_{\mathrm{e}}(\mathcal{X})$ to $\mathcal{P}_{\mathrm{e}}(\mathcal{P}_{\mathrm{e}}(\mathbb{R}))$ via the (non-injective) mapping

$$\mathbb{X} \mapsto \mu_{\mathbb{X}} := \frac{1}{N}\sum_{i=1}^{N} \delta_{d_{\mathcal{X}}(x_i,\cdot)_\sharp \mu}, \quad \delta_{d_{\mathcal{X}}(x_i,\cdot)_\sharp \mu} = \frac{1}{N}\sum_{j=1}^{N} \delta_{d_{\mathcal{X}}(x_i,x_j)}.$$

Now, we can represent two mm-spaces as $\mu_{\mathbb{X}}, \nu_{\mathbb{Y}} \in \mathcal{P}_{\mathrm{e}}(\mathcal{P}_{\mathrm{e}}(\mathbb{R}))$ and compare them using the 1d WoW distance (Mémoli, 2011, 'Third Lower Bound' (TLB')), another SW-based distance (Piening & Beinert, 2025a, 'Sliced Third Lower Bound' (STLB)) or a so-called energy distance (Sato et al., 2020, 'Anchor Energy' (AE)). Alternatively, we may use our SQW distance.

Repeating experiments from (Piening & Beinert, 2025a), we precompute pairwise distance matrices with respect to our distance, TLB, STLB, AE, and the GW distance. Then, we estimate the accuracy of a k-nearest neighbor (KNN) classification by assigning each test point to the majority class among its three nearest neighbors. We average the classification accuracy over 1000 random 25%/75% training/test splits on four datasets. Based on preprocessing from (Piening & Beinert, 2025a), we employ the '2D Shapes' dataset (Beier et al. (2022), $N = 50$), the 'Animals' dataset (Sumner & Popović (2004), $N = 50$) and the 'FAUST' dataset (Bogo et al., 2014) with 500 ('FAUST-500') and 1000 ('FAUST-1000') vertices per shape. We set the kernel parameter σ to 0.01, $R = 10$ and $S = 100$. We use the same integration grid and projection number for STLB. We display the results in Table 1, where we observe comparable performance across all distances and a runtime advantage of SQW and STLB for the large-scale FAUST-1000 dataset, in particular against the GW distance.

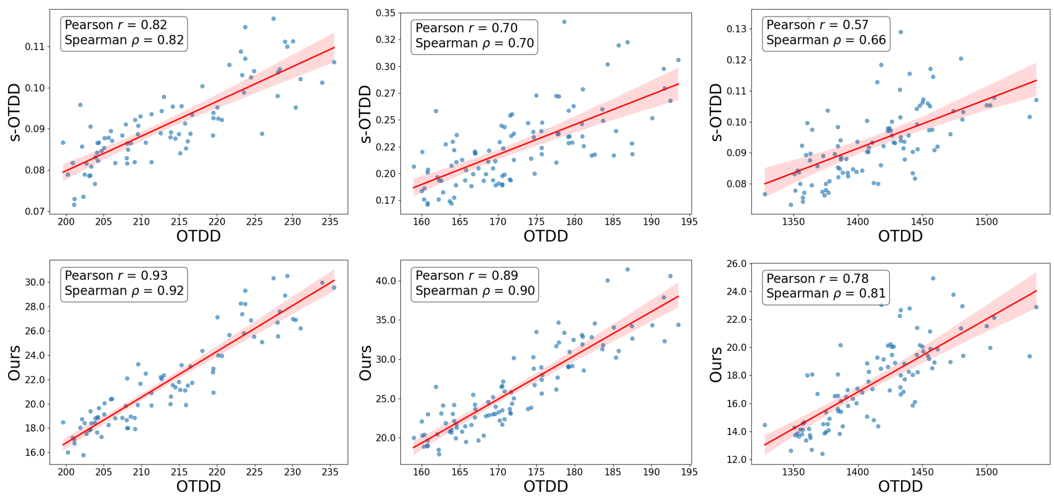
Distance	2D shapes		Animals		FAUST-500		FAUST-1000	
	Acc. (%)	Time (ms)	Acc. (%)	Time (ms)	Acc. (%)	Time (ms)	Acc. (%)	Time (ms)
Ours	99.5±1.2	1.6	99.1±1.3	1.4	38.6±5.7	13.0	42.7 ± 5.9	17.9
TLB	100.0±0.3	0.7	100.0±0.0	0.7	36.7±5.6	22.4	40.2 ± 6.0	86.3
STLB	99.5±1.2	1.6	99.3±1.8	1.2	37.6±5.6	13.0	39.4±5.6	17.8
AE	99.7±0.9	0.7	97.8±1.8	0.6	37.7±5.6	8.0	41.8±5.3	30.7
GW	99.7±0.6	2.7	100.0±0.0	4.9	29.2±4.4	42.0	33.0±5.3	187.8

Table 1: Shape classification with KNN: Mean accuracy (Acc., ↑) and runtime (Time).

5.2 Optimal Transport Dataset Distance

Next, we consider a comparison with the OTDD (Alvarez-Melis & Fusi, 2020) and the s-OTDD (Nguyen et al., 2025). These metrics have been developed to quantify the similarities between labelled datasets in a model-agnostic manner. Such similarity metrics are especially important for applications in transfer learning. In this area, it has been shown empirically that the OTDD and s-OTDD display a strong correlation with the performance gap in transfer learning and classification accuracy in data augmentation, see (Alvarez-Melis & Fusi, 2020). To assess the suitability of our sliced metric as a drop-in replacement for the computationally costly OTDD, we repeat an experiment from (Nguyen et al., 2025). We randomly split MNIST (LeCun et al., 1998), FashionMNIST (Xiao et al., 2017), and CIFAR10 (Krizhevsky, 2009) to create subdataset pairs, each ranging in size from 500 to 1000, and compute the OTDD, the s-OTDD, and our DSW between subdataset pairs.

We plot the results of our distance and the s-OTDD against OTDD for 100 dataset splits in Figure 1, where we include the Pearson and the Spearman correlation coefficients between both sliced distances and the OTDD. As our metric is originally defined on $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ and OTDD and s-OTDD on $\mathcal{P}(\mathcal{Y} \times \mathcal{P}(\mathbb{R}^d))$, we compute OTDD and s-OTDD with the label metric on $\mathcal{Y}$ set to zero for comparability, effectively representing each dataset as an empirial measure in $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ and computing the OTDD via WoW. To be precise, each class-conditional distribution is modeled as $\mu_i \in \mathcal{P}_{\mathrm{e}}(\mathbb{R}^d)$ and the distribution over class-conditional distributions becomes our meta-measure $\mu \in \mathcal{P}_{\mathrm{e}}(\mathcal{P}_{\mathrm{e}}(\mathbb{R}^d))$. Similar to (Nguyen et al., 2025), we estimate both DSW ($R = 10$, $\sigma = 0.1$) and the s-OTDD (with Radon features) with $S = 10,000$ projections. Based on this setting, we employ the original default implementation for the s-OTDD and the 'exact' OTDD. Treating OTDD as our ground truth, we clearly see a stronger correlation between DSW and OTDD for all datasets.



(a) MNIST (b) FashionMNIST (c) CIFAR-10

Figure 1: Scatter plots and correlations (↑) between the s-OTDD and the OTDD (top) and our DSW and the OTDD (bottom) for MNIST (1a), FashionMNIST (1b), and CIFAR-10 (1c).

5.3 Comparing Distributions of Point Clouds

For our next experiment, we assess the potential of DSW for evaluating point cloud generative models, which aim to generate 3D shapes such as chairs or planes. Evaluating such models is challenging since common quality metrics are insensitive to mode collapse (e.g., 'coverage') or tolerate low-quality samples (e.g., 'minimum matching distances'); see Yang et al. (2019) for details. A common remedy is the OT nearest-neighbor accuracy ('OT-NNA') test, which uses 1-nearest-neighbor classification based on pairwise Wasserstein distances between real and generated point clouds. However, for N real and M generated shapes, this requires $(N + M - 1)^2/2$ OT computations, without defining a true metric.

As a natural alternative, one might instead represent batches of real and generated point clouds $\mu_i \in \mathcal{P}_{\mathrm{e}}(\mathbb{R}^3)$ as empirical meta-measures in $\mathcal{P}_{\mathrm{e}}(\mathcal{P}_{\mathrm{e}}(\mathbb{R}^3))$ and compute the WoW, respectively, the DSW distance between a real and a generated batch. To assess the suitability of our resulting quality metric, we consider shapes from ModelNet-10 (Wu et al., 2015) and construct meta-measures

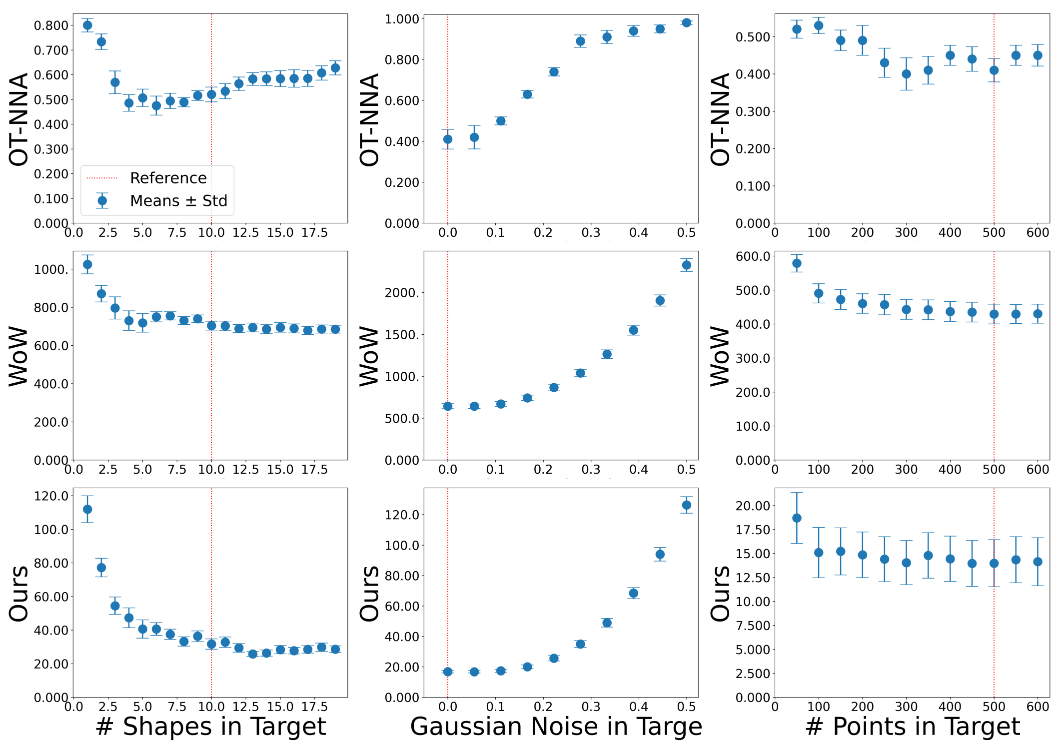
$$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\mu_i} \in \mathcal{P}_{\mathrm{e}}(\mathcal{P}_{\mathrm{e}}(\mathbb{R}^3)), \quad \boldsymbol{\nu} = \frac{1}{M} \sum_{j=1}^{M} \delta_{\nu_j} \in \mathcal{P}_{\mathrm{e}}(\mathcal{P}_{\mathrm{e}}(\mathbb{R}^3)),$$

where the support points are again Euclidean empirical measures

$$\mu_i = \frac{1}{n} \sum_{k=1}^{n} \delta_{x_{i,k}} \in \mathcal{P}_{\mathrm{e}}(\mathbb{R}^3), \quad \nu_j = \frac{1}{m} \sum_{\ell=1}^{m} \delta_{x'_{j,\ell},+\varepsilon} \in \mathcal{P}_{\mathrm{e}}(\mathbb{R}^3) \quad \epsilon \sim \mathcal{N}(0, \sigma_{\text{Noise}}^2 \mathbf{I}_3).$$

Given a certain shape class, e.g., 'chair', we initialize $\boldsymbol{\mu}$ as our fixed reference meta-measure with (downsampled) shapes from the ModelNet-10 training set and $\boldsymbol{\nu}$ as our varying target meta-measure with shapes from the ModelNet-10 test set. To compare OT-NNA, WoW, and our DSW metric, we independently vary the number of target shapes M, the level of Gaussian noise σ_{Noise}, and the point cloud discretization m while fixing the remaining parameters according to the reference $\boldsymbol{\mu}$ (Default parameters: $N = M$, $\sigma_{\text{Noise}} = 0$, $m = n$). The results of our experiment are visualized in Figure 2, where we display the average result of 5 runs with varying M, σ_{Noise}, and n, and the reference ('ground truth') parameter of $\boldsymbol{\mu}$ is marked with a dotted red line. Lower is better for all metrics.

Looking at the number of target shapes M in Subfigure 2a (Class: 'bed', $N = 10$, $\sigma_{\text{Noise}} = 0$, $n = m = 50$), we see that all metrics successfully capture mode collapse, i.e., $M = 1$, and decrease for a larger number of target shapes M. Unlike the plateauing WoW and DSW metrics, OT-NNA displays an undesired behavior by increasing for $M \geq N = 10$, however. As for the random Gaussian perturbations in Subfigure 2b (Class: 'sofa', $N = M = 10$, $n = m = 50$), all three metrics increase with increasing noise. Whereas OT-NNA is more sensitive to small noise levels, the WoW and DSW metrics are more sensitive to high noise levels. Considering the point cloud resolution m in Subfigure 2c (Class: 'monitor', $N = M = 10$, $n = 500$), OT-NNA seems inconsistent regarding the resolution. In contrast, the WoW and DSW metrics are higher for $m \leq 100$ and plateau after a seemingly sufficient resolution has been reached. Overall, we see that WoW and DSW display similar behavior as the OT-NNA and offer the advantage of being unbounded metrics. Additionally, DSW ($S = 10,000$, $R = 50$, $\sigma = 0.1$) offers computational advantages as it takes around 0.25 seconds for $M = N = 10$ and $m = n = 500$, where WoW takes about 4.5 seconds and OT-NNA about 8.5 seconds (on our CPU). This makes it especially suitable for high-resolution point clouds and large point cloud batches.



(a) Number of Target Shapes (b) Noise of Target Points (c) Number of Target Points

Figure 2: OT-NNA, WoW, and our DSW between target and reference point cloud batches for varying numbers of shapes M in target batch (2a), for Gaussian noise σ_{Noise} for target points (2b) and varying target point cloud resolution m (2c). Fixed reference values are marked in red.

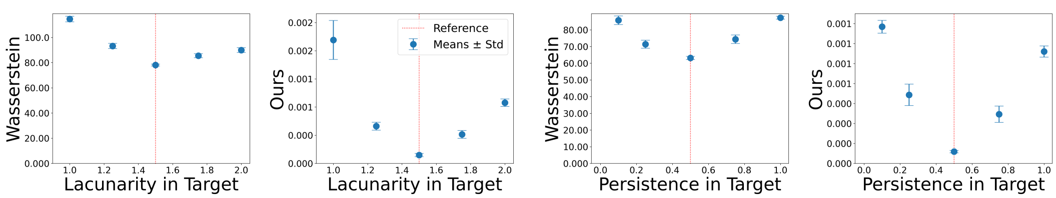
5.4 Comparing Image Distributions via their Patch Distributions

Given the importance of OT for imaging, we conclude with an imaging experiment. Interestingly, OT is utilized on two levels in this area. On the one hand, it is used to compare *two batches of* *images* using pairwise Euclidean distances (Genevay et al., 2018). On the other hand, using OT as a distance *between two individual images* remains of relevance due to the disadvantages of Euclidean distances. Those methods represent images as 2D histograms (Beier et al., 2023; Geuter et al., 2025) or as *patch* distributions (Hertrich et al., 2022; Elnekave & Weiss, 2022; Flotho et al., 2025). Thus, a natural combination is the comparison of image batches using WoW (Dukler et al., 2019).

Since patch-based OT distances serve as a perceptual metric between images (He et al., 2024), we incorporate patch-based image representations into the WoW framework. This is based on parametrizing images via their distribution of localized features. More concretely, we map each (grayscale) image to the empirical measure over all contained (overlapping) square-shaped image regions of size $p \times p$, see (Piening et al., 2024) for an in-depth description. Thus, we may represent each image as an empirical measure $\mu_i \in \mathcal{P}_{\mathrm{e}}(\mathbb{R}^{p^2})$ supported on vectorized patches and a batch of images as an empirical meta-measure $\boldsymbol{\mu} \in \mathcal{P}_{\mathrm{e}}(\mathcal{P}_{\mathrm{e}}(\mathbb{R}^{p^2}))$. In this experiment, we establish a quantitative comparison of image batches using our hierarchical OT framework as an alternative to comparing two image batches using standard OT or a neural-network-based perceptual metric, such as the 'Kernel Inception Distance' (Sutherland et al., 2018, 'KID').

To validate this approach, we consider synthetic 64×64 texture images based on random Perlin noise (Perlin, 1985). This texture synthesis model is controlled by several parameters, among them the *lacunarity* and the *persistence*. Similar to our previous experiment, we initialize a reference meta-measure $\boldsymbol{\mu}$ over images represented as patch distributions according to some reference parameters and compare it to a target meta-measure $\boldsymbol{\nu}$ with varying lacunarity, respectively, persistence. For batch size 32 and patch size $p = 8$, both meta-measures are initialized according to 32 random images, where each image is represented by $(64 - 7) \times (64 - 7) = 3249$ uniformly weighted patches of dimension $8 \times 8 = 64$. In Figure 3, we compare the average behavior over five runs of the standard Wasserstein distance with Euclidean cost between our reference and target images and our patch-based DSW distance ($\sigma = 0.1$, $S = 10,000$ projections, $R = 10$), where the 'ground truth' lacunarity and persistence reference parameters are again marked in red.

We observe that both distances are minimized for the 'true' reference parameters. Still, our patch-based approach is more sensitive to parameter variation and better at discriminating between different batches, see also supplementary comparisons. Note that patch-based WoW calculation takes about 40 seconds, whereas our sliced distance merely requires about one second.



(a) Varying Lacunarity in Perlin Noise (b) Varying Persistence in Perlin Noise

Figure 3: Comparing synthetic texture image batches via Euclidean Wasserstein and our sliced patch-based distance based on varying 'lacunarity' (3a) and 'persistence' (3b). Both distances are minimized for 'true' parameters (red), but our sliced distance leads to clearer discrimination.

6 Conclusion

We introduce a general sliced OT framework for measures on arbitrary Banach spaces. Leveraging the isometry between 1d Wasserstein and $L_2([0, 1])$, Gaussian-process–parametrized L_2-projections, and classical spherical slicing, we define the DSW distance between meta-measures, a well-posed, computationally efficient substitute for WoW. We prove that DSW minimization corresponds to WoW minimization for discretized meta-measures and demonstrate practical effectiveness on datasets, shapes, and images. On the practical side, future work could align DSW with the original OTDD by employing hybrid slicing (Nguyen & Ho, 2024) to extend DSW to $\mathcal{P}_2(\mathcal{Y} \times \mathcal{P}_2(\mathbb{R}^d))$ or integrate convolutional projections (Nguyen & Ho, 2022) similar to the s-OTDD. Also, one might employ our Banach slicing for infinite-dimensional generative models (Hagemann et al., 2025). On the theoretical side, it would be of interest to analyze further topological properties.

References

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21428–21439. Curran Associates, 2020.

Lazar Atanackovic, Xi Zhang, Brandon Amos, Mathieu Blanchette, Leo J Lee, Yoshua Bengio, Alexander Tong, and Kirill Neklyudov. Meta flow matching: Integrating vector fields on the Wasserstein manifold. In *Proceedings of the ICLR'25*. OpenReview.net, 2025.

Matthias Beckmann, Robert Beinert, and Jonas Bresch. Normalized Radon cumulative distribution transforms for invariance and robustness in optimal transport based image classification, 2025. arXiv:2506.08761.

Florian Beier, Robert Beinert, and Gabriele Steidl. On a linear Gromov–Wasserstein distance. *IEEE Transactions on Image Processing*, 31:7292–7305, 2022.

Florian Beier, Johannes von Lindheim, Sebastian Neumayer, and Gabriele Steidl. Unbalanced multi-marginal optimal transport. *Journal of Mathematical Imaging and Vision*, 65(3):394–413, 2023.

Vladimir Igorevich Bogachev. *Gaussian measures*. Number 62. American Mathematical Soc., 1998.

Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings of CVPR'14*. IEEE, 2014.

Clément Bonet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, and Minh Tan Pham. Spherical sliced-Wasserstein. In *Proceedings of the ICLR'23*. OpenReview.net, 2023a.

Clément Bonet, Laetitia Chapel, Lucas Drumetz, and Nicolas Courty. Hyperbolic sliced-Wasserstein via geodesic and horospherical projections. In *Proceedings of the Topological, Algebraic and Geometric Learning Workshops 2023*, pp. 334–370, 2023b.

Clément Bonet, Benoıt Malézieux, Alain Rakotomamonjy, Lucas Drumetz, Thomas Moreau, Matthieu Kowalski, and Nicolas Courty. Sliced-Wasserstein on symmetric positive definite matrices for M/EEG signals. In *Proceedings of the ICML'23*, pp. 2777–2805. PMLR, 2023c.

Clément Bonet, Lucas Drumetz, and Nicolas Courty. Sliced-Wasserstein distances and flows on Cartan-Hadamardmanifolds. *Journal of Machine Learning Research*, 26(32):1–76, 2025a.

Clément Bonet, Christophe Vauthier, and Anna Korba. Flowing datasets with Wasserstein over Wasserstein gradient flows. In *Proceedings of ICML'25*. OpenReview.net, 2025b.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Nicolas Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, Université Paris Sud–Paris XI, Orsay, France, 2013. PhD Thesis.

Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2018.

Yukun Chen, Jianbo Ye, and Jia Li. Aggregated Wasserstein distance and state registration for hidden Markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9): 2133–2147, 2019.

Nathaël Da Costa, Marvin Pförtner, Lancelot Da Costa, and Philipp Hennig. Sample path regularity of Gaussian processes from the covariance kernel. *arXiv preprint arXiv:2312.14886*, 2023.

Juan Antonio Cuesta-Albertos, Ricardo Fraiman, and Thomas Ransford. A sharp form of the cramér–wold theorem. *Journal of Theoretical Probability*, 20(2):201–209, 2007.

Julie Delon and Agnes Desolneux. A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

Yonatan Dukler, Wuchen Li, Alex Lin, and Guido Montúfar. Wasserstein of Wasserstein loss for learning generative models. In *Proceedings of ICML'19*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1716–1725. PMLR, 2019.

Mourad El Hamri, Younes Bennani, and Issam Falih. Hierarchical optimal transport for unsupervised domain adaptation. *Machine Learning*, 111(11):4159–4182, 2022.

Ariel Elnekave and Yair Weiss. Generating natural images with direct patch distributions matching. In *Proceedings of the ECCV'22*, pp. 544–560. Springer, 2022.

Ryszard Engelking. General topology. *Revised and completed edition*, 1989.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proceedings of the AISTATS'19*, pp. 2681–2690. PMLR, 2019.

Philipp Flotho, Moritz Piening, Anna Kukleva, and Gabriele Steidl. T-FAKE: Synthesizing thermal images for facial landmarking. In *Proceedings of the CVPR'25*, pp. 26356–26366, 2025.

Robert Garrett, Trevor Harris, Zhuo Wang, and Bo Li. Validating climate models with spherical convolutional Wasserstein distance. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59119–59149. Curran Associates, 2024.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of the AISTATS'18*, pp. 1608–1617. PMLR, 2018.

Jonathan Geuter, Gregor Kornhardt, Ingimar Tomasson, and Vaios Laschos. Universal neural optimal transport. In *Proceedings of the ICML'25*. OpenReview.net, 2025.

Paul Hagemann, Sophie Mildenberger, Lars Ruthotto, Gabriele Steidl, and Nicole Tianjiao Yang. Multilevel diffusion: Infinite dimensional score-based diffusion models for image generation. *SIAM Journal on Mathematics of Data Science*, 7(3):1337–1366, 2025.

Ruiyu Han. Sliced Wasserstein distance between probability measures on Hilbert spaces. *arXiv preprint arXiv:2307.05802*, 2023.

Doron Haviv, Aram-Alexandre Pooladian, Dana Pe'er, and Brandon Amos. Wasserstein flow matching: Generative modeling over families of distributions. In *Proceedings of the ICML'25*. OpenReview.net, 2025.

Jiaqi He, Zhihua Wang, Leon Wang, Tsein-I Liu, Yuming Fang, Qilin Sun, and Kede Ma. Multiscale sliced Wasserstein distances as perceptual color difference measures. In *Proceedings of the ECCV'24*, pp. 425–442. Springer, 2024.

A Heppes. On the determination of probability distributions of more dimensions by their projections. *Acta Mathematica Hungarica*, 7(3-4):403–410, 1956.

Johannes Hertrich, Antoine Houdard, and Claudia Redenbach. Wasserstein patch prior for image superresolution. *IEEE Transactions on Computational Imaging*, 8:693–704, 2022.

Johannes Hertrich, Tim Jahn, and Michael Quellmalz. Fast summation of radial kernels via QMC slicing. In *Proceedings of the ICLR'25*. OpenReview.net, 2025.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pp. 2278–2324. IEEE, 1998.

Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Springer, Berlin, 1991.

John Lee, Max Dabagia, Eva Dyer, and Christopher Rozell. Hierarchical optimal transport for multimodal distribution alignment. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, 2019.

Lin Lin, Wei Shi, Jianbo Ye, and Jia Li. Multisource single-cell data integration by MAW barycenter for Gaussian mixture models. *Biometrics*, 79(2):866–877, 2023.

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11:417–487, 2011.

Karim Nadjahi, François-Xavier Vialard, and Gabriel Peyré. Fast approximation of the Sliced-Wasserstein distance using concentration of random projections. In *Proceedings of the ICML'21*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7946–7956. PMLR, 2021.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15657–15669. Curran Associates, 2020.

Khai Nguyen and Nhat Ho. Revisiting sliced Wasserstein on images: From vectorization to convolution. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17788–17801. Curran Associates, 2022.

Khai Nguyen and Nhat Ho. Energy-based sliced Wasserstein distance. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18046–18075. Curran Associates, 2023.

Khai Nguyen and Nhat Ho. Hierarchical hybrid sliced Wasserstein: A scalable metric for heterogeneous joint distributions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 108140–108166. Curran Associates, 2024.

Khai Nguyen and Peter Mueller. Summarizing Bayesian nonparametric mixture posterior–sliced optimal transport metrics for Gaussian mixtures. *arXiv preprint arXiv:2411.14674*, 2024.

Khai Nguyen, Nicola Bariletto, and Nhat Ho. Quasi-Monte Carlo for 3d sliced Wasserstein. In *Proceedings of the ICLR'24*. OpenReview.net, 2024.

Khai Nguyen, Hai Nguyen, Tuan Pham, and Nhat Ho. Lightspeed geometric dataset distance via sliced optimal transport. In *Proceedings of ICML'25*. OpenReview.net, 2025.

Trung Nguyen, Quang-Hieu Pham, Tam Le, Tung Pham, Nhat Ho, and Binh-Son Hua. Point-set distances for learning representations of 3d point clouds. In *Proceedings of the CVPR'21*, pp. 10478–10487. IEEE, 2021.

Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: with applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Moritz Piening and Robert Beinert. A novel sliced fused Gromov-Wasserstein distance. *arXiv preprint arXiv:2508.02364*, 2025a.

Moritz Piening and Robert Beinert. Slicing the Gaussian mixture Wasserstein distance. *Transactions on Machine Learning Research*, 2025b.

Moritz Piening, Fabian Altekrüger, Johannes Hertrich, Paul Hagemann, Andrea Walther, and Gabriele Steidl. Learning from small data sets: Patch-based regularizers in inverse problems for image reconstruction. *GAMM-Mitteilungen*, 47(4):e202470002, 2024.

Michael Quellmalz, Robert Beinert, and Gabriele Steidl. Sliced optimal transport on the sphere. *Inverse Problems*, 39(10):105005, 2023.

Michael Quellmalz, Léo Buecher, and Gabriele Steidl. Parallelly sliced optimal transport on spheres and on the rotation group. *Journal of Mathematical Imaging and Vision*, 66(6):951–976, 2024.

Balram S Rajput and Stamatis Cambanis. Gaussian processes and gaussian measures. *The Annals of Mathematical Statistics*, pp. 1944–1952, 1972.

Ryoma Sato, Marco Cuturi, Makoto Yamada, and Hisashi Kashima. Fast and robust comparison of probability measures in heterogeneous spaces, 2020. arXiv:2002.01615.

Bernhard Schmitzer and Christoph Schnörr. A hierarchical approach to optimal transport. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 452–464. Springer, 2013.

Konrad Schmüdgen et al. *The moment problem*, volume 9. Springer, 2017.

Jan Stanczuk, Christian Etmann, Lisa Maria Kreusser, and Carola-Bibiane Schönlieb. Wasserstein GANs work because they fail (to approximate the Wasserstein distance). *arXiv preprint arXiv:2103.01678*, 2021.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.

Simon Steuernagel, Aaron Kurda, and Marcus Baum. Point cloud registration based on Gaussian mixtures and pairwise Wasserstein distances. In *Proceedings of the SDF-MFI'23*, pp. 1–8. IEEE, 2023.

Robert W. Sumner and Jovan Popović. Mesh data from deformation transfer for triangle meshes. Available online at % urlhttp://people.csail.mit.edu/sumner/research/deftransfer/data.html, 2004. Accessed: 2025-06-03.

JD Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *Proceedings of the ICLR'18*, volume 6, 2018.

Eloi Tanguy, Rémi Flamary, and Julie Delon. Reconstructing discrete measures from projections. consequences on the empirical sliced Wasserstein distance. *Comptes Rendus. Mathématique*, 362 (G10):1121–1129, 2024.

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Aad W Van Der Vaart, J Harry Van Zanten, et al. Reproducing kernel hilbert spaces of Gaussian priors. *IMS Collections*, 3:200–222, 2008.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, Providence, RI, 2003.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the CVPR'25*. IEEE, 2015.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017. Available via arXiv:1708.07747.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the ICCV'19*, pp. 4541–4550. IEEE, 2019.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. Hierarchical optimal transport for document representation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, 2019.

Hanna Ziesche and Leonel Rozo. Wasserstein gradient flows for optimizing Gaussian mixture policies. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21058–21080. Curran Associates, 2023.

A NON-SPHERICAL SLICED WASSERSTEIN DISTANCE ON BANACH SPACES

Here, we present proofs for Section 3.1. For clarity, we restate and prove our statements from the main paper as smaller statements.

Lemma A.1. *For $\mu, \nu \in \mathcal{P}_2(U)$, $\theta \in U^* \mapsto \mathrm{W}(\pi_{\theta,\sharp}\,\mu, \pi_{\theta,\sharp}\,\nu;\, \mathbb{R})$ is Lipschitz continuous.*

Proof. We generalize the proof of (Han, 2023, Lem. 2.3). For this, let $\theta_1, \theta_2 \in U^*$ be arbitrary. Using the triangle inequality and its reverse, we have

$$\begin{aligned}
&|\mathrm{W}(\pi_{\theta_1,\sharp}\,\mu, \pi_{\theta_1,\sharp}\,\nu;\, \mathbb{R}) - \mathrm{W}(\pi_{\theta_2,\sharp}\,\mu, \pi_{\theta_2,\sharp}\,\nu;\, \mathbb{R})| \\
&\leq \left[\mathrm{W}(\pi_{\theta_1,\sharp}\,\mu, \pi_{\theta_2,\sharp}\,\mu;\, \mathbb{R}) + \mathrm{W}(\pi_{\theta_2,\sharp}\,\mu, \pi_{\theta_1,\sharp}\,\nu;\, \mathbb{R})\right] \\
&\qquad - \left[\mathrm{W}(\pi_{\theta_2,\sharp}\,\mu, \pi_{\theta_1,\sharp}\,\nu;\, \mathbb{R}) - \mathrm{W}(\pi_{\theta_1,\sharp}\,\nu, \pi_{\theta_2,\sharp}\,\nu;\, \mathbb{R})\right] \\
&= \mathrm{W}(\pi_{\theta_1,\sharp}\,\mu, \pi_{\theta_2,\sharp}\,\mu;\, \mathbb{R}) + \mathrm{W}(\pi_{\theta_1,\sharp}\,\nu, \pi_{\theta_2,\sharp}\,\nu;\, \mathbb{R}).
\end{aligned}$$

For the first term on the left-hand side, it follows

$$\begin{aligned}
\mathrm{W}^2(\pi_{\theta_1,\sharp}\,\mu, \pi_{\theta_2,\sharp}\,\mu;\, \mathbb{R}) &\leq \int_{\mathbb{R}^2} |t_1 - t_2|^2 \,\mathrm{d}(\pi_{\theta_1}, \pi_{\theta_2})_\sharp\,\mu(t_1, t_2) = \int_U |\langle x, \theta_1 - \theta_2\rangle|^2 \,\mathrm{d}\mu(x) \\
&\leq \|\theta_1 - \theta_2\|_{U^*}^2 \int_U \|x\|_U^2 \,\mathrm{d}\mu(x) = \|\theta_1 - \theta_2\|_{U^*}^2 M_2(\mu),
\end{aligned}$$

where $M_2(\mu) \coloneqq \int_U \|x\|_U^2 \,\mathrm{d}\mu(x)$ is the second moment of μ. Using an analogous estimate for the second term, we obtain

$$|\mathrm{W}(\pi_{\theta_1,\sharp}\,\mu, \pi_{\theta_1,\sharp}\,\nu;\, \mathbb{R}) - \mathrm{W}(\pi_{\theta_2,\sharp}\,\mu, \pi_{\theta_2,\sharp}\,\nu;\, \mathbb{R})| \leq \|\theta_1 - \theta_2\|_{U^*} \left(M_2^{1/2}(\mu) + M_2^{1/2}(\nu)\right). \qquad \square$$

This allows us to prove the first part of Theorem 3.1.

Proposition A.2. *For $\xi \in \mathcal{P}_2(U^*)$, the ξ-based SW distance is well-defined.*

Proof. The Lipschitz continuity in Lemma A.1 implies that the integrand in the formulation of the ξ-based SW (4) is measurable. To show that $\mathrm{SW}(\mu, \nu;\, \xi)$ is finite for $\mu, \nu \in \mathcal{P}_2(U)$, let $\gamma \in \Gamma(\mu, \nu)$ realize $\mathrm{W}(\mu, \nu;\, U)$. Because of

$$\begin{aligned}
\mathrm{W}^2(\pi_{\theta,\sharp}\,\mu, \pi_{\theta,\sharp}\,\nu;\, \mathbb{R}) &\leq \int_{U \times U} |\langle x_1, \theta\rangle - \langle x_2, \theta\rangle|^2 \,\mathrm{d}\gamma(x_1, x_2) \\
&\leq \|\theta\|_{U^*}^2 \int_{\mathcal{X} \times \mathcal{X}} \|x_1 - x_2\|_U^2 \,\mathrm{d}\gamma(x_1, x_2) = \|\theta\|_{U^*}^2\, \mathrm{W}^2(\mu, \nu; U),
\end{aligned} \tag{8}$$

the ξ-based SW distance is bounded by $\mathrm{SW}(\mu, \nu;\, \xi) \leq \mathrm{W}(\mu, \nu; U)\, M_2^{1/2}(\xi)$. $\square$

Now, we prove the second part of Theorem 3.1.

Theorem A.3. *Let $\xi \in \mathcal{P}_2(U^*)$ be such that $\operatorname{supp}\xi \cap \operatorname{span}\theta \notin \{\emptyset, \{0\}\}$ for any $\theta \in U^*$, then $\mathrm{SW}(\cdot, \cdot;\, \xi)$ defines a metric on $\mathcal{P}_2(U)$. Otherwise, $\mathrm{SW}(\cdot, \cdot; \xi)$ defines at least a pseudo-metric.*

Proof. Positivity, symmetry, and triangle inequality follow from the corresponding properties of the Wasserstein distance. For the definiteness, assume that $\mathrm{SW}(\mu, \nu;\, \xi) = 0$, which implies $\mathrm{W}(\pi_{\theta,\sharp}\,\mu, \pi_{\theta,\sharp}\,\nu;\, \mathbb{R}) = 0$ for almost all $\theta \in U^*$ with respect to ξ. The Lipschitz continuity in Lemma A.1 implies $\mathrm{W}(\pi_{\theta,\sharp}\,\mu, \pi_{\theta,\sharp}\,\nu;\, \mathbb{R}) = 0$ and thus $\pi_{\theta,\sharp}\,\mu = \pi_{\theta,\sharp}\,\nu$ for all $\theta \in \operatorname{supp}\xi$. Now, let $\theta' \in U^*$ be arbitrary. By assumption, we find $t \in \mathbb{R} \setminus \{0\}$ such that $t\theta' \in \operatorname{supp}\xi$. Furthermore, we have

$$\begin{aligned}
\int_U \mathrm{e}^{\mathrm{i}\langle x, \theta'\rangle} \,\mathrm{d}\mu(x) &= \int_U \mathrm{e}^{\mathrm{i}\langle x, t\theta'\rangle \frac{1}{t}} \,\mathrm{d}\mu(x) = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}s\frac{1}{t}} \,\mathrm{d}\pi_{t\theta',\sharp}\,\mu(s) \\
&= \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}s\frac{1}{t}} \,\mathrm{d}\pi_{t\theta',\sharp}\,\nu(s) = \int_U \mathrm{e}^{\mathrm{i}\langle x, \theta'\rangle} \,\mathrm{d}\nu(x).
\end{aligned}$$

Since every measure on U has a unique characteristic function, see (Ledoux & Talagrand, 1991, § 2.1), we conclude $\mu = \nu$. $\square$

We continue with Proposition 3.2.

Proposition A.4. *Let $\xi_1, \xi_2 \in \mathcal{P}_2(U^*)$ be equivalent. If $\mathrm{d}\xi_1/\,\mathrm{d}\xi_2$ and $\mathrm{d}\xi_2/\,\mathrm{d}\xi_1$ are bounded, then we find $c_1, c_2 > 0$ such that*

$$c_1\,\mathrm{SW}(\mu,\nu;\,\xi_1) \le \mathrm{SW}(\mu,\nu;\,\xi_2) \le c_2\,\mathrm{SW}(\mu,\nu;\,\xi_1) \quad \forall \mu,\nu \in \mathcal{P}_2(U).$$

Proof. Exploiting the bounded Radon–Nikodým derivatives, we obtain

$$\mathrm{SW}^2(\mu,\nu;\,\xi_2) = \int_{U^*} \mathrm{W}^2(\pi_{\theta,\sharp}\,\mu, \pi_{\theta,\sharp}\,\nu;\,\mathbb{R})\,\tfrac{\mathrm{d}\xi_2}{\mathrm{d}\xi_1}(\theta)\,\mathrm{d}\xi_1(\theta) \le \left\|\tfrac{\mathrm{d}\xi_2}{\mathrm{d}\xi_1}(\theta)\right\|_{L^\infty_{\xi_1}(U^*)}\,\mathrm{SW}^2(\mu,\nu;\,\xi_1).$$

□

We employ this to prove Proposition 3.3.

Proposition A.5. *For $\xi \in \mathcal{P}(\mathbb{R}^d)$ absolutely continuous and $\mathrm{d}\xi/\,\mathrm{d}\eta$, $\mathrm{d}\eta/\,\mathrm{d}\xi$ bounded, for $\eta \sim \mathcal{N}(0, \mathbf{I}_d)$, there exist $c_1, c_2 > 0$ such that*

$$c_1\,\mathrm{SW}(\mu,\nu\,;\,\xi) \le \mathrm{SW}(\mu,\nu) \le c_2\,\mathrm{SW}(\mu,\nu;\,\xi) \quad \forall \mu,\nu \in \mathcal{P}_2(\mathbb{R}^d).$$

Proof. This follows directly from Proposition A.4 and the identity $\mathrm{SW}(\mu,\nu) = C_d\,\mathrm{SW}(\mu,\nu;\eta)$, see (Nadjahi et al., 2021, Prop. 1). □

We extend our findings from the main paper with a statement about the computational approximation via Monte Carlo. In particular, the ξ-based SW distance can be numerically approximated by

$$\widehat{\mathrm{SW}}^2(\mu,\nu;\,\xi) \approx \frac{1}{S}\sum_{s=1}^{S} \mathrm{W}^2(\pi_{\theta_s,\sharp}\,\mu, \pi_{\theta_s,\sharp}\,\nu;\,\mathbb{R}), \quad \theta_s \sim \xi \text{ iid}.$$

It is well-known that such Monte Carlo estimates have a convergence rate of $\mathcal{O}(1/\sqrt{S})$ for S random projections (Nadjahi et al., 2020). Given suitable conditions, a similiar results holds for this estimate.

Proposition A.6. *For $\xi \in \mathcal{P}_4(U^*)$, it holds*

$$\mathbb{E}_{\theta_1,\ldots,\theta_S}|\widehat{\mathrm{SW}}^2(\mu,\nu;\,\xi) - \mathrm{SW}^2(\mu,\nu;\,\xi)| \le \frac{1}{\sqrt{S}}\,\mathrm{std}_\theta\,\mathrm{W}^2(\pi_{\theta,\sharp}\,\mu, \pi_{\theta,\sharp}\,\nu;\,\mathbb{R}).$$

Proof. Using Hölder's inequality, and exploiting that the directions θ_s are independent and identically distributed, we have

$$\begin{aligned}
&\mathbb{E}_{\theta_1,\ldots,\theta_S}|\widehat{\mathrm{SW}}^2(\mu,\nu;\,\xi) - \mathrm{SW}^2(\mu,\nu;\,\xi)| \\
&\le \Big(\int_{U^*}\cdots\int_{U^*}\Big|\frac{1}{S}\sum_{s=1}^{S}\mathrm{W}^2(\pi_{\theta_s,\sharp}\,\mu, \pi_{\theta_s,\sharp}\,\nu;\,\mathbb{R}) - \mathrm{SW}^2(\mu,\nu;\,\xi)\Big|^2\,\mathrm{d}\xi(\theta_1)\cdots\mathrm{d}\xi(\theta_S)\Big)^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{S}}\Big(\sum_{s=1}^{S}\int_{U^*}|\mathrm{W}^2(\pi_{\theta_s,\sharp}\,\mu, \pi_{\theta_s,\sharp}\,\mu;\,\mathbb{R}) - \mathrm{SW}^2(\mu,\nu;\,\xi)|^2\,\mathrm{d}\xi(\theta_s)\Big)^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{S}}\,\mathrm{std}_\theta\,\mathrm{W}^2(\pi_{\theta,\sharp}\,\mu, \pi_{\theta,\sharp}\,\nu;\,\mathbb{R}),
\end{aligned}$$

where the standard deviation exists due to $\mathrm{W}^4(\pi_{\theta,\sharp}\,\mu, \pi_{\theta,\sharp}\,\nu;\,\mathbb{R}) \le \|\theta\|_{U^*}^4\,\mathrm{W}^4(\mu,\nu;U)$, cf. (8). □

B Double-Sliced Wasserstein Distance

B.1 Metric Properties

To prove the positive definiteness of our double-sliced metric for empirical meta-measures, we utilize an extension of the 'Cramer–Wold' theorem by Cuesta-Albertos et al. (2007) about the set of projections required to separate measures on $\mathbb{R}^d$. The statement is based on the so-called *Carleman condition*. A measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ fulfills this condition if all moments

$$M_p(\mu) := \int \|x\|^p \,\mathrm{d}\mu, \quad p \geq 1,$$

are finite and it holds that

$$\sum_{p=1}^{\infty} M_p^{-1/p} = \infty.$$

This condition is fulfilled for compactly supported measures (Schmüdgen et al., 2017, Ch. 14) and, in particular, empirical measures. We also refer to (Heppes, 1956; Tanguy et al., 2024) for similar results targeted at empirical measures.

Lemma B.1. ***(Cuesta-Albertos et al., 2007, Corr. 3.2)*** *Given measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ that fulfill the Carleman condition and a set $S \subset \mathbb{S}^{d-1}$ of positive surface measure with*

$$\mathrm{W}(\pi_{\theta\sharp}\mu, \pi_{\theta\sharp}\nu; \mathbb{R}) = 0 \quad \textit{for all } \theta \in S, \tag{9}$$

it holds that $\mu = \nu$.

This allows us to prove the metric properties presented in Theorem 3.1.

Proposition B.2. *Given a positive $\xi \in \mathcal{P}_2(L_2(Y))$, $\mathbf{DSW}$ defines a metric on $\mathcal{P}_\mathrm{e}(\mathcal{P}_\mathrm{e}(\mathbb{R}^d))$.*

Proof. 'Pseudo-metric': The symmetry and triangle inequality are trivial and follow directly from the ambient spaces of the embedded measures and the properties of the Wasserstein distance.

'Positive Definiteness': We aim to prove that

$$\mathbf{DSW}(\boldsymbol{\mu}, \boldsymbol{\nu};\, \xi) = 0 \iff \boldsymbol{\mu} = \boldsymbol{\nu}.$$

for empirical meta-measures. Therefore, assume that $\mathbf{DSW}(\boldsymbol{\mu}, \boldsymbol{\nu};\, \xi) = 0$ for $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_\mathrm{e}(\mathcal{P}_\mathrm{e}(\mathbb{R}^d))$.

Due to $\mathbf{DSW}(\boldsymbol{\mu}, \boldsymbol{\nu};\, \xi) = 0$, we know that $\mathbf{SQW}(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu}, \boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\nu};\, \xi) = 0$ for all $\theta \in \mathbb{S}^{d-1}$ except for a zero measure set Z. Since $\mathbf{SQW}$ is a metric on $\mathcal{P}_2(\mathbb{R})$, we thus know that $\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu} = \boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\nu}$ for every $\theta \in \mathbb{S}^{d-1} \setminus Z$. Now, this means that there exists a $\boldsymbol{\gamma}_\theta^* \in \Gamma(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu}, \boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\nu}) \subset \mathbb{R}_{\geq 0}^{n \times m}$, such that

$$\langle \boldsymbol{\gamma}_\theta^*, C_\theta \rangle = 0, \tag{10}$$

where $C_\theta = (\mathrm{W}^2(\pi_{\theta,\sharp}\mu_i, \pi_{\theta,\sharp}\nu_j))_{i,j} \subset \mathbb{R}_{\geq 0}^{n \times m}$ and $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. Here, our costs and transport plans take matrix form due to the empirical measure structure.

We now want to find a suitable transport plan from this set of transport plans to construct an upper bound for the meta-measure metric $\mathbf{W}(\boldsymbol{\mu}, \boldsymbol{\nu}; \mathbb{R}^d)$. Consider the set of index pairs $\mathrm{Idx} = \{(i,j) \,|\, \mu_i \neq \nu_j\}$. If this set is empty, then we are done. Otherwise, we know from Lemma B.1 that there exists no set $S \subset \mathbb{S}^{d-1}$ of positive measure for (μ_i, ν_j), $(i,j) \in \mathrm{Idx}$, such that (9) is fulfilled for all $\theta \in S$. Conversely, for $(i,j) \in \mathrm{Idx}$, it holds that $(C_\theta)_{i,j} > 0$ for every θ outside a zero measure set $Z_{i,j}$. Subsequently, it has to hold that $(\gamma_\theta^*)_{i,j} = 0$ outside the zero measure set $Z \cup Z_{i,j}$ due to (10).

Now, we have $(\gamma_\theta^*)_{i,j} = 0$ or $C_{i,j} = \mathrm{W}^2(\mu_i, \nu_j) = 0$ for some $\theta \in \mathbb{S}^{d-1}$ outside the zero measure set $(Z \cup (\cup_{(i,j) \in \mathrm{Idx}} Z_{i,j}))$. Thus, it holds for almost every θ that

$$\langle \boldsymbol{\gamma}_\theta^*, C \rangle = 0.$$

Since this expression is an upper bound of $\mathbf{W}^2(\boldsymbol{\mu}, \boldsymbol{\nu}; \mathbb{R}^d)$, it follows that $\mathbf{W}(\boldsymbol{\mu}, \boldsymbol{\nu}; \mathbb{R}^d) = 0$. This concludes the proof since the Wasserstein distance is a positive definite metric. ☐

Remark. Although we stated our statement for the empirical measures employed in our experiments, our proof merely requires that all measures satisfy the Carleman condition. As an example, this would be fulfilled for mixtures of compactly supported measures, i.e., for $\mathcal{P}_{\mathrm{e}}(\mathcal{P}_2(\mathcal{X}))$, $\mathcal{X} \subset \mathbb{R}^d$ compact. Note that $\mathcal{P}_{\mathrm{e}}(\mathcal{P}_2(\mathcal{X}))$ is dense in $\mathcal{P}_{\mathrm{e}}(\mathcal{P}_2(\mathcal{X}))$ (Villani, 2003, Ch. 6). Moreover, Lemma A.1 allows us to show the Lipschitz continuity of DSW on $\mathcal{P}_2(\mathcal{P}_2(\mathcal{X}))$. Combining all of this with statements on continuous extensions of metrics on topological spaces Engelking (1989), we can expect DSW to be a metric on $\mathcal{P}_2(\mathcal{P}_2(\mathcal{X}))$.

B.2 Relationships between Metric

In this section, we aim to prove our convergence result. We point out that (Han, 2023, Thm. 3.4) contains a proof of weak convergence for measures on general Hilbert spaces. However, the underlying argument appears to rely on an application of a convergence result in infinite-dimensional settings whose validity in this context is, to the best of our understanding, not fully clear.

To prove our convergence statement, we separate the proof into a couple of smaller statements. As a first step, we show that our sliced metrics produce lower bounds for $\mathbf{W}$. We continue with a lemma that relates the subset of $\mathcal{P}_2(L_2([0,1]))$ supported on piecewise constant step functions to Euclidean measures. Based on this, we prove a statement about the equivalence between $\mathbf{DSW}$ and $\mathbf{SW}$. Lastly, we prove a statement about the equivalence between $\mathbf{SW}$ and $\mathbf{W}$, which utilizes the compactness of the support.

First, we state a proposition that allows us to bound $\mathbf{W}$ from below via $\mathbf{DSW}$ and $\mathbf{SW}$.

Proposition B.3. *It holds* $C_\xi\, \mathbf{DSW}(\boldsymbol{\mu}, \boldsymbol{\nu};\, \xi) \leq \mathbf{SW}(\boldsymbol{\mu}, \boldsymbol{\nu};\, \xi) \leq \mathbf{W}(\boldsymbol{\mu}, \boldsymbol{\nu};\, \mathbb{R}^d)$ *for* $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ *and* $\xi \in \mathcal{P}_2(L_2([0,1]))$.

Proof. DSW is essentially a double integral. For $v \in L_2([0,1])$, $\theta \in \mathbb{S}^{d-1}$ and $\tilde{\boldsymbol{\gamma}} \in \Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$, the DSW integrand can be estimated using the Cauchy-Schwarz inequality by

$$
\begin{aligned}
\mathrm{W}^2(\pi_{v,\sharp}(q_\sharp(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu})), \pi_{v,\sharp}(q_\sharp(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\nu});\, \mathbb{R}) &\leq \int_{\mathcal{P}_2(\mathbb{R}^d)\times\mathcal{P}_2(\mathbb{R}^d)} |\langle q(\pi_{\theta,\sharp}\mu) - q(\pi_{\theta,\sharp}\nu), v\rangle|^2 \,\mathrm{d}\tilde{\boldsymbol{\gamma}}(\mu,\nu) \\
&\leq \int_{\mathcal{P}_2(\mathbb{R}^d)\times\mathcal{P}_2(\mathbb{R}^d)} \|q(\pi_{\theta,\sharp}\mu) - q(\pi_{\theta,\sharp}\nu)\|^2 \|v\|^2 \,\mathrm{d}\tilde{\boldsymbol{\gamma}}(\mu,\nu).
\end{aligned}
$$

Subsequently, it holds

$$
\mathrm{W}^2(\pi_{v,\sharp}(q_\sharp(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu})), \pi_{v,\sharp}(q_\sharp(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\nu});\, \mathbb{R}) \leq \|v\|^2 \mathbf{W}_2^2(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu}, \boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\nu};\, \mathbb{R}) \leq \|v\|^2 \mathbf{W}_2^2(\boldsymbol{\mu}, \boldsymbol{\nu};\, \mathbb{R}^d).
$$

The last inequality follows from $\mathrm{W}(\pi_{\theta,\sharp}\mu, \pi_{\theta,\sharp}\nu;\, \mathbb{R}) \leq \mathrm{W}(\mu, \nu;\, \mathbb{R}^d)$ for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\theta \in \mathbb{S}^{d-1}$. Because of $\xi \in \mathcal{P}_2(L_2([0,1]))$, we know that $\int_{L_2([0,1])} \|v\|^2 \,\mathrm{d}\xi(v) = M_2(\xi) < \infty$. Integration with respect to ξ and the uniform measure on $\mathbb{S}^{d-1}$ gives the statement with $C_\xi = 1/\sqrt{M_2(\xi)}$. □

From this relation, we can easily see that $\mathbf{W}(\boldsymbol{\mu}_n, \boldsymbol{\mu};\, \mathbb{R}^d) \to 0$ results in $\mathbf{SW}(\boldsymbol{\mu}_n, \boldsymbol{\mu};\, \xi) \to 0$ and that $\mathbf{SW}(\boldsymbol{\mu}_n, \boldsymbol{\mu};\, \xi) \to 0$ results in $\mathbf{DSW}(\boldsymbol{\mu}_n, \boldsymbol{\mu};\, \xi) \to 0$.

In the following statement, we use indicator functions $\mathbf{1}_{x\in S}$ that take the value 1 for $x \in S$ and 0 otherwise.

Lemma B.4. *For* $\mu_n, \mu \in \mathcal{P}_2(L_2([0,1]))$ *only supported on fixed-length step functions, i.e., sums of indicator functions of the form*

$$
S_P = \left\{ \sum_{i=1}^{\tilde{n}} f_i \mathbf{1}_{x\in P_i} \;:\; f_i \in \mathbb{R} \right\}
$$

for a fixed partition $\dot\bigcup_{i=1}^{\tilde{n}} P_i = [0,1]$, *and a positive Gaussian measure* $\xi \in \mathcal{P}_2(L_2([0,1]))$, *we have*

$$
\mathrm{SW}(\mu_n, \mu;\, \xi) \to 0 \iff \mathrm{W}(\mu_n, \mu;\, L_2([0,1]) \to 0.
$$

Proof. To prove the statement, we want to leverage the established metric equivalence between the Wasserstein and the classical sliced Wasserstein distance on $\mathcal{P}_2(\mathbb{R}^{\tilde{n}})$. Therefore, we construct an isometric mapping $\mathrm{IM}_P : S_P \to \mathbb{R}^{\tilde{n}}$ similar to (Piening & Beinert, 2025a). Instead of considering probability measures on the infinite-dimensional space $L_2([0,1])$, this allows us to consider Euclidean probability measures. We define our S_P-isometric mapping on the space $L_2([0,1])$ via

$$\mathrm{IM}_P : L_2([0,1]) \to \mathbb{R}^{\tilde{n}},$$
$$f \mapsto \left(|P_i|^{\frac{1}{2}} \int_{P_i} f(x)\, dx \right)_{i=1}^{\tilde{n}} \in \mathbb{R}^{\tilde{n}}.$$

This is an isometry on S_p because for $f^{(1)}, f^{(2)} \in S_P$ it holds that

$$\|f^{(1)} - f^{(2)}\|^2 = \sum_{i=1}^{\tilde{n}} |P_i|(f_i^{(1)} - g_i)^2 = \|\mathrm{IM}(f^{(1)}) - \mathrm{IM}(f^{(2)})\|^2.$$

Now, this means that

$$\mathrm{W}^2(\mu_n, \mu;\, L_2([0,1])) = \mathrm{W}^2(\mathrm{IM}_{P,\sharp}\, \mu_n, \mathrm{IM}_{P,\sharp}\, \mu;\, \mathbb{R}^{\tilde{n}}). \tag{11}$$

Now, we aim to construct a similar relation for the ξ-based sliced Wasserstein distance. We require $\xi_P \in \mathcal{P}_2(\mathbb{R}^{\tilde{n}})$ based on $\xi \in \mathcal{P}_2(L_2([0,1]))$. Therefore, we again employ our mapping to link the L_2-projection to a projection on $\mathbb{R}^{\tilde{n}}$. For $f \in S_P$, $g \in L_2([0,1])$, we have $\langle f, g\rangle = \langle \mathrm{IM}_P(f), \mathrm{IM}_P(g)\rangle$. Hence, we define $\xi_P \coloneqq \mathrm{IM}_{P,\sharp}\, \xi$. As a sliced counterpart of (11), we get that

$$\mathrm{SW}^2(\mu_n, \mu; \xi) = \mathrm{SW}^2(\mathrm{IM}_\sharp\, \mu_n, \mathrm{IM}_\sharp\, \mu; \xi_P).$$

Now, we want to utilize Proposition 3.3 to connect our ξ-based sliced Wasserstein to the classical sliced Wasserstein distance. Indeed, since ξ is a Gaussian measure and IM_P is linear, ξ_P is Gaussian by definition of a Gaussian measure. Moreover, it is a nondegenerate Gaussian since ξ is positive. Thus, $\mathrm{SW}(\mathrm{IM}_\sharp\, \mu_n, \mathrm{IM}_\sharp\, \mu; \xi_P)$ is topologically equivalent to the classical sliced distance $\mathrm{SW}(\mathrm{IM}_\sharp\, \mu_n, \mathrm{IM}_\sharp\, \mu)$ by Proposition 3.3. Since SW and W induce the same weak topology on $\mathcal{P}_2(\mathbb{R}^{\tilde{n}})$ (Nadjahi et al., 2020), we overall conclude that

$$\begin{aligned} & \mathrm{SW}(\mu_n, \mu;\, \xi) \to 0 \\ \iff & \mathrm{SW}(\mathrm{IM}_\sharp\, \mu_n, \mathrm{IM}_\sharp\, \mu; \xi_P) \to 0 \\ \iff & \mathrm{SW}(\mathrm{IM}_\sharp\, \mu_n, \mathrm{IM}_\sharp\, \mu) \to 0 \\ \iff & \mathrm{W}(\mathrm{IM}_\sharp\, \mu_n, \mathrm{IM}_\sharp\, \mu;\, \mathbb{R}^{\tilde{n}}) \to 0 \\ \iff & \mathrm{W}(\mu_n, \mu;\, L_2([0,1])) \to 0. \end{aligned}$$

□

Proposition B.5. *Given a positive Gaussian $\xi \in \mathcal{P}_2(L_2([0,1]))$ and empirical meta-measures $\boldsymbol{\mu}_n, \boldsymbol{\mu} \in \mathcal{P}_\mathrm{e}^N(\mathcal{P}^{\tilde{n}}(\mathcal{X}))$, $\mathcal{X} \subset \mathbb{R}^d$ compact, we have that*

$$\mathbf{DSW}(\boldsymbol{\mu}_n, \boldsymbol{\mu};\, \xi) \to 0 \iff \mathbf{SW}(\boldsymbol{\mu}_n, \boldsymbol{\mu};\, \xi) \to 0.$$

Proof. We assume that

$$\mathbf{DSW}^2(\boldsymbol{\mu}, \boldsymbol{\nu};\, \xi) = \int_{\mathbb{S}^{d-1}} \mathrm{SW}^2\left(q_\sharp(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu}_n), q_\sharp(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu});\, \xi\right) \mathrm{d}\mathbb{S}^{d-1}(\theta) \to 0.$$

It follows that

$$\mathrm{SW}(q_\sharp \boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu}_n, q_\sharp \boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu};\, \xi) \to 0$$

for almost any $\theta \in \mathbb{S}^{d-1}$. Since we are dealing with fixed, uniform weights, all quantile functions are piecewise constant step functions with a fixed step length. By Lemma B.4, it thus follows that

$$\mathrm{W}(q_\sharp \boldsymbol{\pi}_{\theta,\sharp}\mu_n, q_\sharp \boldsymbol{\pi}_{\theta,\sharp}\mu_n;\, L_2([0,1]) = \mathbf{W}(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu}_n, \boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu};\, \xi) \to 0$$

for almost every $\theta \in \mathbb{S}^{d-1}$. Since the compact support results in boundedness, it thus follows from the dominated convergence theorem that

$$\mathbf{SW}^2(\boldsymbol{\mu}_n, \boldsymbol{\mu};\, \xi) = \int_{\mathbb{S}^{d-1}} \mathbf{W}^2(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu}, \boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\nu};\, \mathbb{R})\, \mathrm{d}\mathbb{S}^{d-1}(\theta) \to 0.$$

□

Now, we state the last result. Note that our proof is inspired by a proof in (Piening & Beinert, 2025b).

Proposition B.6. *Given $\boldsymbol{\mu}_n, \boldsymbol{\mu} \in \mathcal{P}_{\mathrm{e}}^N(\mathcal{P}_{\mathrm{e}}^{\tilde{n}}(\mathcal{X}))$, $\mathcal{X} \subset \mathbb{R}^d$ compact, it holds that*

$$\mathbf{SW}(\boldsymbol{\mu}_n, \boldsymbol{\mu};\, \xi) \to 0 \iff \mathbf{W}(\boldsymbol{\mu}_n, \boldsymbol{\mu};\, \mathbb{R}^d) \to 0.$$

Proof. We write

$$\boldsymbol{\mu}_n = \frac{1}{N}\sum_{i=1}^N \delta_{\mu_{n,i}}, \quad \boldsymbol{\mu} = \frac{1}{N}\sum_{j=1}^N \delta_{\mu_j}.$$

Now, we assume that

$$\mathbf{SW}^2(\boldsymbol{\mu}_n, \boldsymbol{\mu};\, \xi) = \int_{\mathbb{S}^{d-1}} \sum_{k=1}^N \sum_{\ell=1}^N \boldsymbol{\gamma}_{\theta,k,\ell}^{n,*} \mathbf{W}_2^2(\pi_{\theta,\sharp}\, \mu_{n,k}, \pi_{\theta,\sharp}\, \mu_\ell;\, \mathbb{R})\, \mathrm{d}\mathbb{S}^{d-1}(\theta) \to 0,$$

where $\boldsymbol{\gamma}_\theta^{n,*} \in \Gamma(\boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu}_n, \boldsymbol{\pi}_{\theta,\sharp}\boldsymbol{\mu}) \subset \mathbb{R}_{\geq 0}^{N\times N}$ denotes the optimal projected WoW plan for a fixed $\theta \in \mathbb{S}^{d-1}$. Because of $L_2(\mathbb{S}^{d-1})$ convergence, for any subsequence of $\boldsymbol{\mu}_n$, we find a further subsequence $(\boldsymbol{\mu}_{n_m})_{m\in\mathbb{N}}$ such that

$$\sum_{k=1}^N \sum_{\ell=1}^N \boldsymbol{\gamma}_{\theta,k,\ell}^{n_m,*} \mathbf{W}_2^2(\pi_{\theta,\sharp}\, \mu_{n_m,k}, \pi_{\theta,\sharp}\, \mu_\ell;\, \mathbb{R}) \to 0 \quad \text{for almost every } \theta \in \mathbb{S}^{d-1} \tag{12}$$

pointwisely. Since $\mathcal{X}$ is compact, $(\boldsymbol{\mu}_{n_m})_{m\in\mathbb{N}}$ can be chosen such that $\mu_{n_m,k}$ converges to some $\tilde{\mu}_k \in \mathcal{P}_2(\mathbb{R}^d)$. Moreover, since $\boldsymbol{\mu}_n$ and $\boldsymbol{\mu}$ are both empirical meta-measures with N support points, we can assume without loss of generality that $\boldsymbol{\gamma}_{\theta,k,\ell}^{n,*}$ is a permutation matrix, i.e., $\boldsymbol{\gamma}_{\theta,k,\ell}^{n,*} \in \{0, \frac{1}{N}\}^{N\times N}$ with only one positive value per row or column (Peyré & Cuturi, 2019, Prop. 2.1).

For any $\theta \in \mathbb{S}^{d-1}$ such that (12) holds true, it follows that $\boldsymbol{\gamma}_{\theta,k,\ell}^{n_m,*} \to \frac{1}{N}$ in the case of $\boldsymbol{\pi}_{\theta,\sharp}\, \tilde{\boldsymbol{\mu}}_k = \boldsymbol{\pi}_{\theta,\sharp}\, \boldsymbol{\mu}_\ell$ and $\boldsymbol{\gamma}_{\theta,k,\ell}^{n_m,*} \to 0$ otherwise. It follows that either $\boldsymbol{\gamma}_{\theta,k,\ell}^{\theta,*} \to 0$ or $\mathrm{W}^2(\mu_{n_m,k}, \mu_\ell) \to 0$. Due to the compactness assumption, we know that $\boldsymbol{\gamma}_{\theta,k,\ell}^{n_m,*}$ and $\mathrm{W}^2(\mu_{n_m,k}, \mu_\ell)$ are bounded and thus

$$\mathbf{W}^2(\mu_n, \mu;\, \mathbb{R}^d) \leq \sum_{k=1}^K \sum_{\ell=1}^K \boldsymbol{\gamma}_{\theta,k,\ell}^{n_m,*}\, \mathrm{W}^2(\mu_{n_m,k}, \mu_\ell;\, \mathbb{R}^d) \to 0 \qquad \text{as} \qquad n \to \infty,$$

Because this holds true for any subsequence of $(\boldsymbol{\mu}_n)_{n\in\mathbb{N}}$, the statement follows. □

Combining Proposition B.5 and Proposition B.6 gives the second statement of Theorem 4.1.

C Additional Details and Experiments

C.1 Use of LLMs

LLMs have been used to a limited extent to improve grammar and wording in this paper.

C.2 Implementation Details

All experiments were conducted with *Python* on a system equipped with a 13th Gen Intel Core i5-13600K CPU and an NVIDIA GeForce RTX 3060 GPU with 12 GB of memory. For all experiments concerning the s-OTDD[1], OTDD, and STLB[2], we employ the official implementations for algorithms and experiments based on the corresponding public GitHub repositories. Unless stated in the experimental description, we use the default hyperparameters for algorithms, including the entropic regularization parameter for OTDD computation. For all other Euclidean Wasserstein computations, we employ the *geomloss* package (Feydy et al., 2019) to estimate entropically regularized Wasserstein distances, where we set the entropic '*blur*' parameter to 0.01. All WoW distances are calculated by estimating a pairwise Wasserstein cost matrix using this *geomloss* geomloss and finally solving the unregularized OT problem given this cost matrix using the *POT* package.

For our own implementation, we employ linear interpolation between the two closest support points for our quadrature grid points. Moreover, instead of sampling the Gaussian processes v and the unit directions θ completely independently, we sample 10 or 100 random Gaussian processes v for each sampled unit direction θ to reduce the number of quantile computations. For our low-dimensional point cloud experiment and our mid-dimensional patch experiment, we employ 100 random 'outer projections' θ and 100 random 'inner projections' v per θ ($10,000$ in total). For the high-dimensional (s-)OTDD experiments, we employ 1000 random 'outer projections' θ and 10 random 'inner projections' v per θ (also $10,000$ in total).

[1]Code: `https://github.com/hainn2803/s-OTDD`.

[2]Code: `https://github.com/MoePien/slicing_fused_gromov_wasserstein`.

C.3 Sliced Functional Optimal Transport on $L_2([0,1])$

Here, we want to showcase the properties of the ξ-based SW distance on $\mathcal{P}_2(L_2([0,1]))$ by looking at increasingly finer function discretizations for different kernel parameters σ. We consider two empirical measure pair in $\mathcal{P}_2(L_2([0,1]))$ defined via

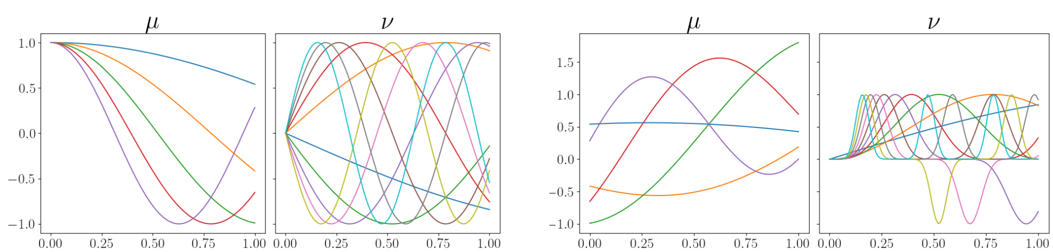
$$f_i^{(1)}(x) = \cos(ix), \quad h_j^{(1)}(x) = \sin(jx + j\pi), \quad f_i^{(2)}(x) = \cos(ix+i) + \sin(x), \quad h_j^{(2)}(x) = \sin(jx)^j$$

and

$$\mu^{(k)} = \frac{1}{5}\sum_{i=1}^{5} \delta_{f_i^{(k)}}, \quad \nu^{(k)} = \frac{1}{10}\sum_{j=1}^{10} \delta_{h_j^{(k)}}, \quad k = 1, 2.$$

In Figure 5, we plot our Monte Carlo estimate for the ξ-based SW between $\mu^{(1)}$ and $\nu^{(1)}$ resp. $\mu^{(2)}$ and $\nu^{(2)}$ estimates for different Gaussian bandwidth parameters σ and equispaced discretization grids with varying size R ($20 \le R \le 100$). To investigate the limit case for $\sigma \to 0$, we also include the case of isotropic Gaussian slicing directions $\theta_l \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$ due to $(k_\sigma(y_i, y_j))_{1 \le i,j \le N} \to \mathbf{I}_N$. While this limit case is well-defined for finite discretization, the limit process would be a white noise process, i.e., a process with sample paths outside of $L_2([0,1])$. On the one hand, we observe that the SW estimate depends heavily on the grid resolution R, especially for isotropic Gaussian directions and for σ small. On the other hand, we see that our numerical estimate is less sensitive to the discretization for larger σ and remains stable given a sufficiently large grid. To get accurate estimates for our comparison, we employ $S = 10{,}000$. We plot all employed functions in Figure 4.

In the main paper, we only employed the radial basis function (RBF) kernel k_σ. However, we point out that we might use other kernels, such as the Brownian motion (BM) kernel $k(s,t) = \min(\{s,t\})$. As an example, we repeat the experiment portrayed in Figure 5 with the BM kernel and plot the results in the last column of Figure 5. Similar to an RBF kernel with high σ, we see that the resulting sliced distance is invariant to the discretization.



$\mu^{(1)}$ and $\nu^{(1)}$. $\mu^{(2)}$ and $\nu^{(2)}$.

Figure 4: Plots of the support functions of the two empirical measure pairs from Section C.3.

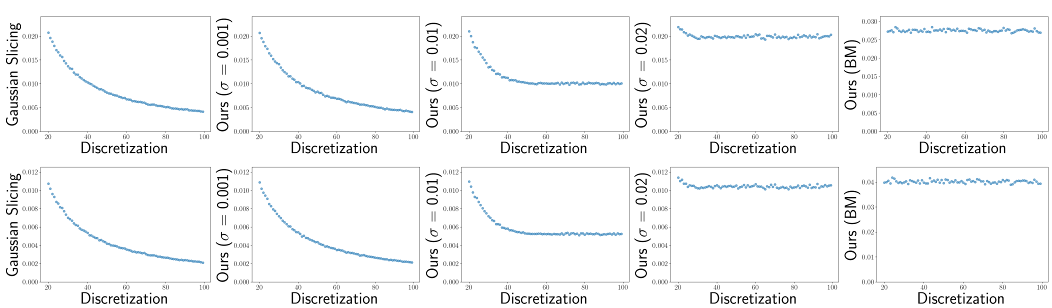


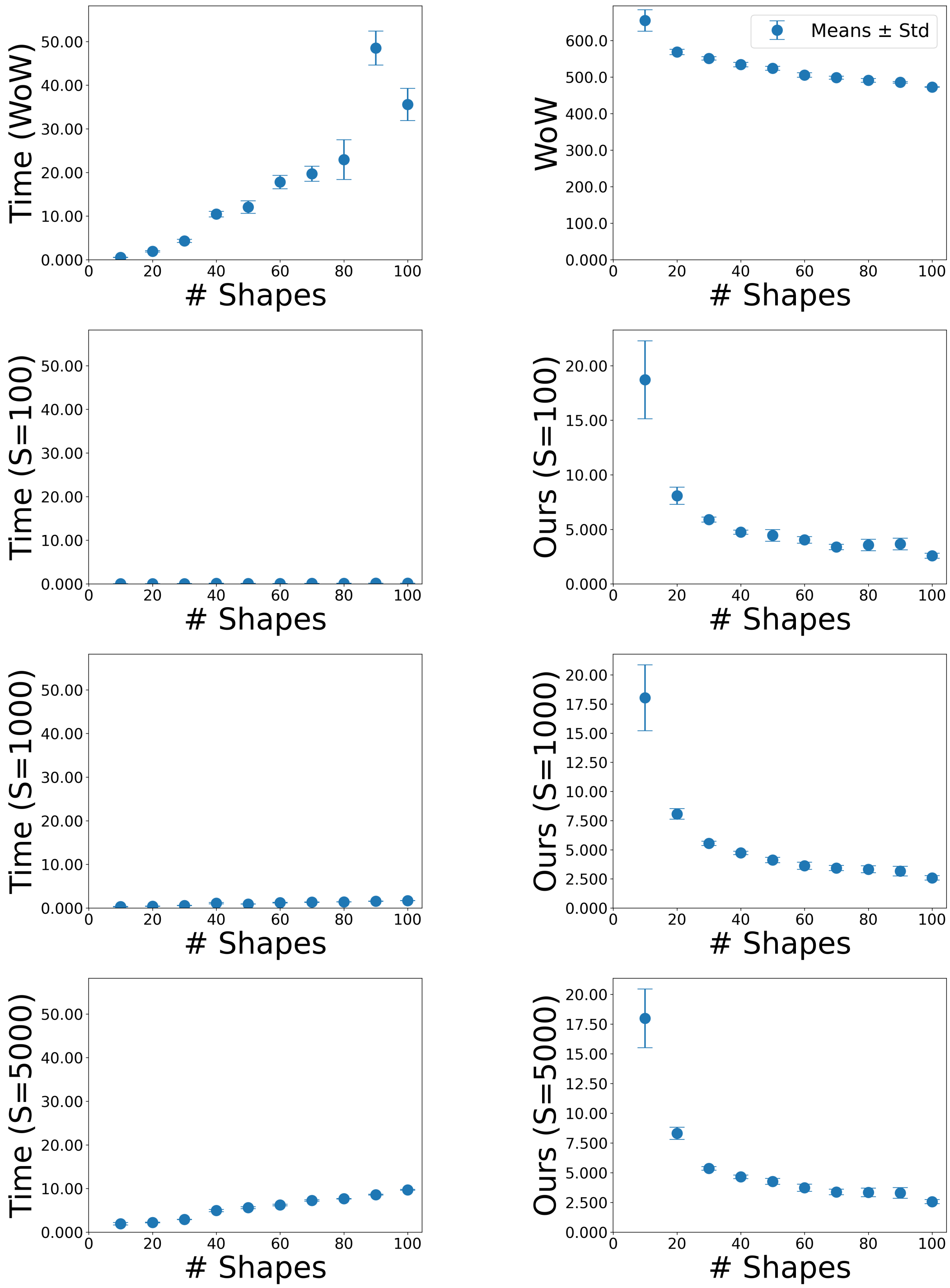
Figure 5: Dependence between function discretization and SW estimates between $\mu^{(1)}$ and $\nu^{(1)}$ (upper row) and $\mu^{(2)}$ and $\nu^{(2)}$ (bottom row) for different kernel choice. Smaller σ values lead to less correlated Gaussian slicing directions for the k_σ kernel, where we include the limit case of fully uncorrelated Gaussian slicing directions in the first column. For such small σ, the numerical estimator depends heavily on the discretization.

C.4 Extension of Section 5.3 on Point Cloud Comparison

We extend the point cloud experiments from Section 5.3 by adding two experiments as an analysis of the projection number and runtime. In particular, we vary only the number of 'inner' or 'outer' projections per experiment, see Section C.2. In our point cloud experiments, the runtime hinges on the number of shapes (N and M) and the discretization of the shapes (n and m). For this analysis, we sample only from the 'chair' class without Gaussian noise and $R = 10$. All results are averaged over five runs.

For our first experiment, we set $n = m = 50$ and vary only $N = M = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100$. For each pair of sampled shape sets, we then compute the WoW distance and our DSW distance. To analyze the impact of the projection number, we calculate it with $S = 100$ (10 outer, 10 inner projections), $S = 1000$ (10 outer, 100 inner p.), and $S = 5000$ (10 outer, 500 inner p.). The results are visualized in Figure 6. Note that we observed a rather high variance for WoW runtime in this experiment, generally. As a result, the plotted WoW runtime estimates in Figure 6 vary rather drastically. Nevertheless, we observe a seemingly polynomial runtime increase for WoW in terms of the number of shapes $N = M$, whereas we only we observe a quasi-linear runtime increase for DSW in terms of $N = M$. As for the projection number, we observe a linear runtime increase in terms of S. Moreover, we observe a (small) reduction in the variance of the distance estimate for higher S.

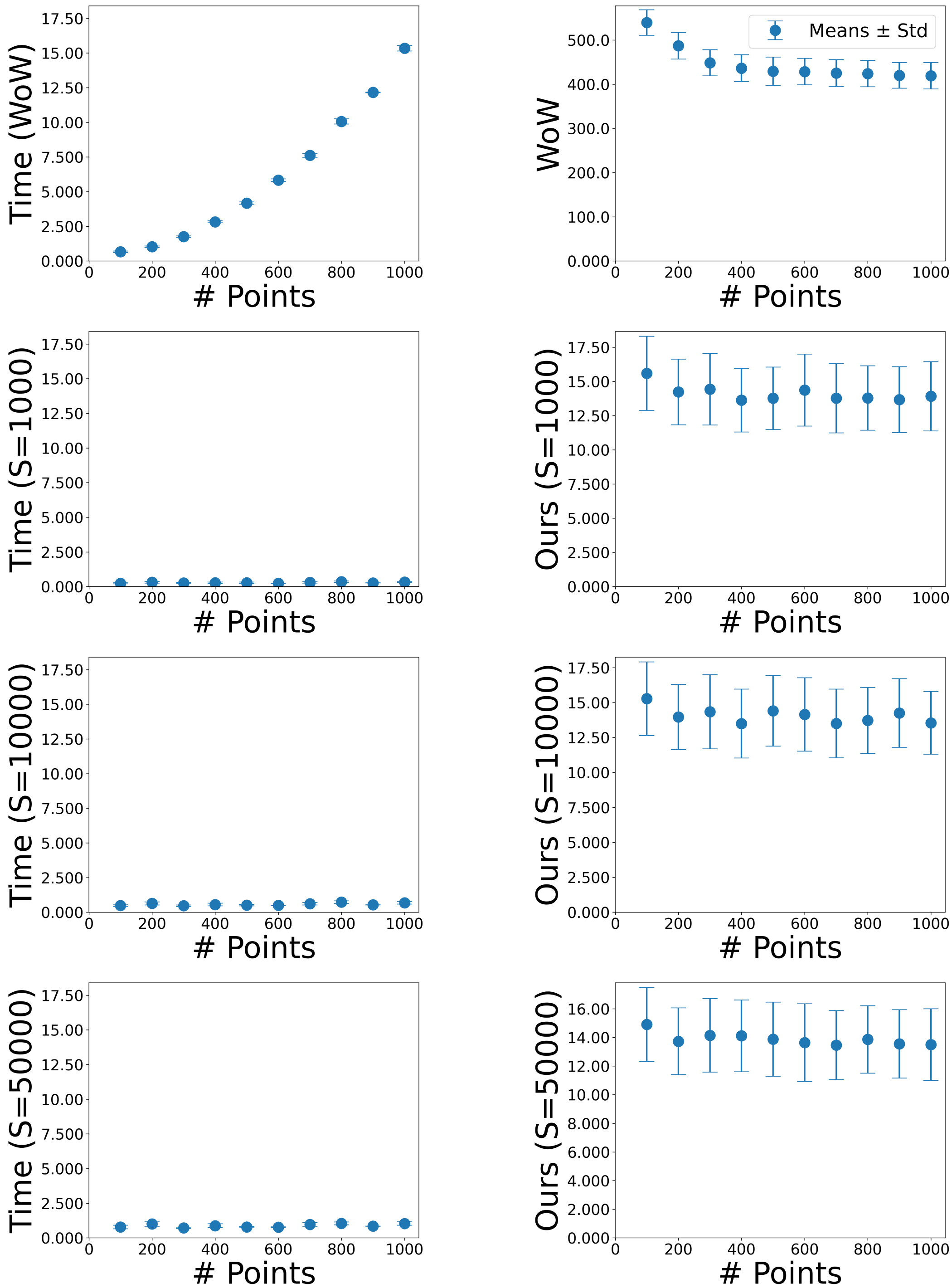
For our second experiment, we set $N = M = 10$ and vary only $n = m = 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000$. Again, we compute WoW and DSW. For this experiment, we calculate DSW with $S = 1000$ (100 outer, 10 inner projections), $S = 10,000$ (100 outer, 100 inner p.), and $S = 50,000$ (100 outer, 500 inner p.). The results are visualized in Figure 7. We observe similar results as before, i.e., polynomial runtime increase for WoW and quasi-linear runtime increase for DSW.



Runtime in seconds.

Distance estimates.

Figure 6: Averaged WoW and DSW estimates between sets of point clouds for 10 to 100 shapes and projection number $S = 100$, $S = 1,000$, $S = 5,000$.



Runtime in seconds.

Distance estimates.

Figure 7: Averaged WoW and DSW estimates between sets of point clouds with 100 to 1000 points per shape and projection number $S = 1000$, $S = 10,000$, $S = 50,000$.

C.5 Extension of Section 5.4 on Image Comparison via Patches

C.5.1 Representing Images via Patches

We formalize the patch extraction. For a grayscale image $\mathrm{Img} \in \mathbb{R}^{h\times w}$ define the patch extractor

$$\mathrm{Patch}_k^p : \mathbb{R}^{h\times w} \to \mathbb{R}^{p^2}, \qquad k = 1, \dots, n_p,$$

with $n_p = (h - p + 1)(w - p + 1)$. Write

$$z_k := \mathrm{Patch}_k^p(\mathrm{Img}) \in \mathbb{R}^{p^2},$$

so that the empirical patch distribution is

$$\mu_{\mathrm{Img}} = \frac{1}{n_p} \sum_{k=1}^{n_p} \delta_{z_k} \in \mathcal{P}_{\mathrm{e}}(\mathbb{R}^{p^2}).$$

Its support is

$$\mathrm{supp}(\mu_{\mathrm{Img}}) = \{\mathrm{Patch}_k^p(\mathrm{Img}) : k = 1, \dots, n_p\} \subset \mathbb{R}^{p^2},$$

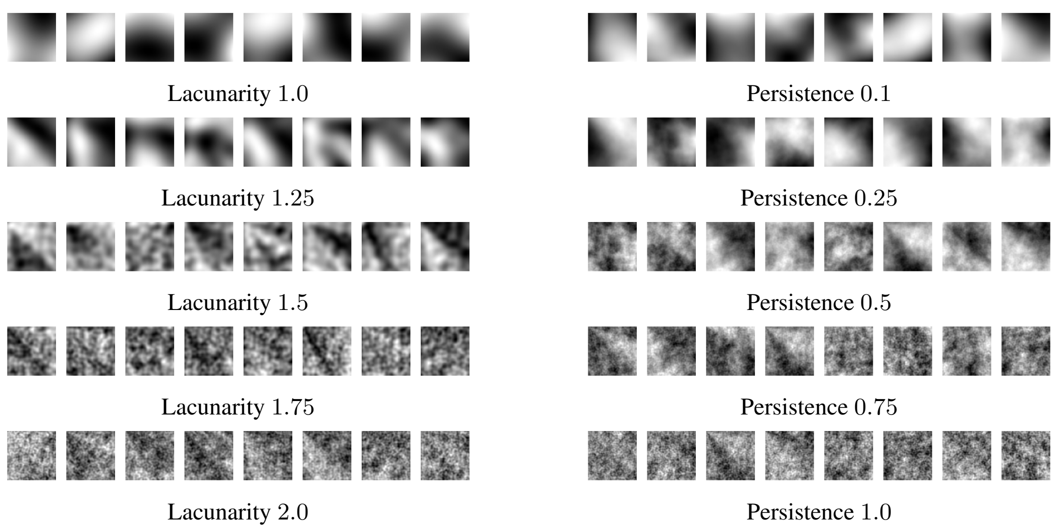
and for a batch $\{\mathrm{Img}_i\}_{i=1}^B$ the meta-measure is

$$\boldsymbol{\mu} = \frac{1}{B} \sum_{i=1}^{B} \delta_{\mu_{\mathrm{Img}_i}} \in \mathcal{P}_{\mathrm{e}}(\mathcal{P}_{\mathrm{e}}(\mathbb{R}^{p^2})).$$

C.5.2 Additional Experimental Details

In the experiment from Section 5.4, we compare distributions over synthetic texture images. We visualize samples from our random Perlin texture model (Perlin, 1985) in Figure 8. Note that our images with varying lacunarity (8a) are all generated with the following Perlin parameters: persistence of 1, scale of 100, 6 octaves. The generation model will be released with the code. For our images with varying persistence (8b), we use different Perlin parameters: lacunarity of 2.5, scale of 100, 5 octaves. Note that while the resulting images in Figure 8a and Figure 8b look rather similar, the ones from Figure 8a display a higher blur and less high-frequency artifacts.

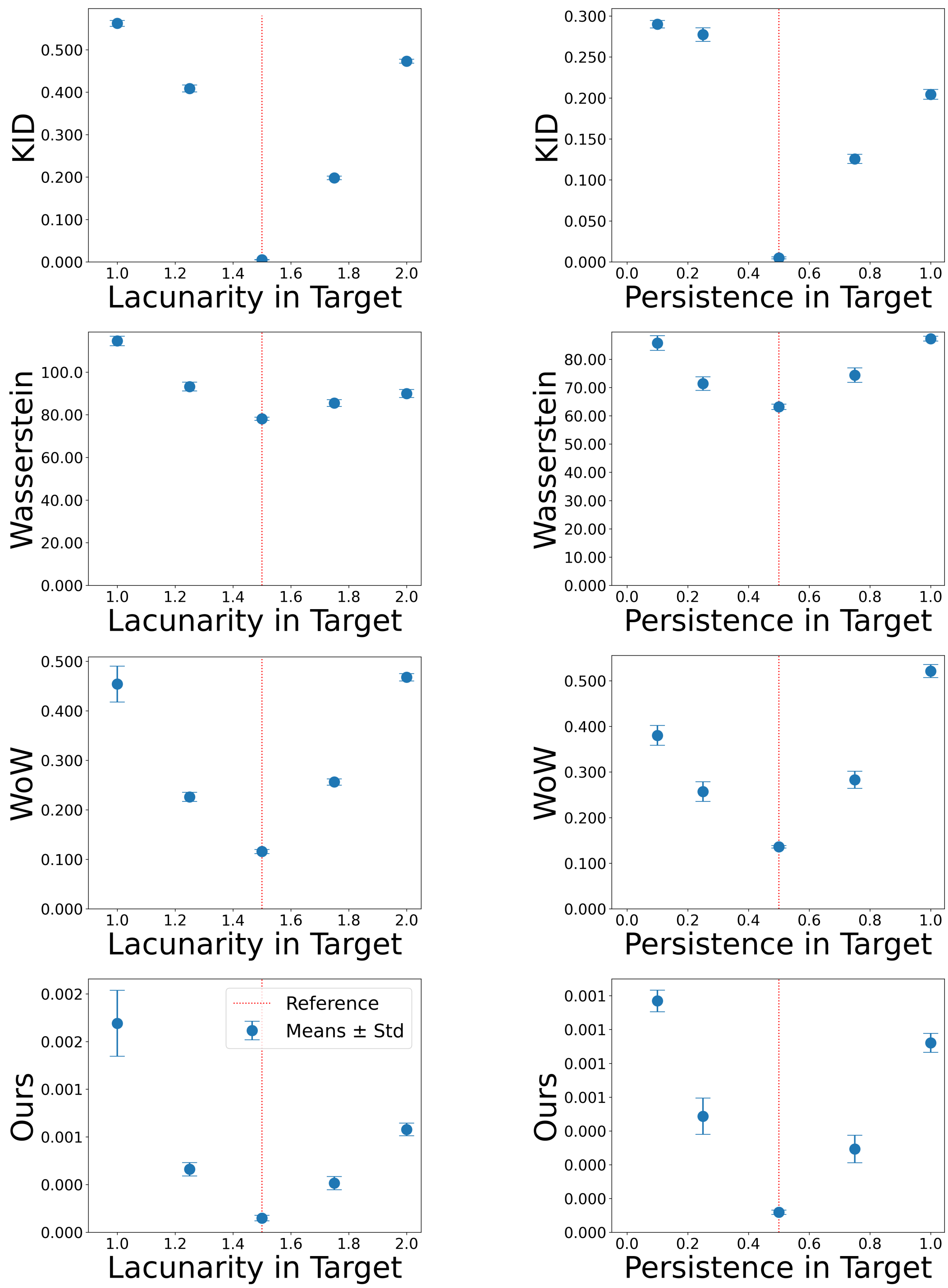
Moreover, we extend Figure 3. In addition to the Wasserstein distance between images represented as Euclidean points and our patch-based DSW distance plotted in the original Figure 3, we present the extended Figure 9 by adding the patch-based WoW distance and the '*Kernel Inception Distance*' (KID) between the distributions of texture images. The patch-based WoW distance is computed on the same patch meta-measures as our patch-based DSW distance. The KID is based on the latent space of a pretrained neural network, see (Sutherland et al., 2018). We see that the DSW and the WoW distance lead to similar results. Also, both are aligned with the KID.



(a) Varying Lacunarity in Perlin Noise.

(b) Varying Lacunarity in Perlin Noise.

Figure 8: Samples from our Perlin texture noise for varying lacunarity (8a) and 'persistence' (8b).



(a) Varying Lacunarity in Perlin Noise

(b) Varying Persistence in Perlin Noise

Figure 9: Comparing synthetic texture image batches via Euclidean Wasserstein, our sliced patch-based distance, patch-based WoW, and the KID for varying 'lacunarity' (3a) and 'persistence' (9b).