NEURAL HAMILTON–JACOBI CHARACTERISTIC FLOWS FOR OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a novel framework for solving optimal transport (OT) problems based on the Hamilton–Jacobi (HJ) equation, whose viscosity solution uniquely characterizes the OT map. By leveraging the method of characteristics, we derive closed-form, bidirectional transport maps, thereby eliminating the need for numerical integration. The proposed method adopts a pure minimization framework: a single neural network is trained with a loss function derived from the method of characteristics of the HJ equation. This design guarantees convergence to the optimal map while eliminating adversarial training stages, thereby substantially reducing computational complexity. Furthermore, the framework naturally extends to a wide class of cost functions and supports class-conditional transport. Extensive experiments on diverse datasets demonstrate the accuracy, scalability, and efficiency of the proposed method, establishing it as a principled and versatile tool for OT applications with provable optimality.

1 INTRODUCTION

Optimal transport (OT) is a fundamental problem that seeks the most cost-efficient transform from one probability distribution into another by minimizing a transportation cost function, which quantifies the effort to move mass. With its strong theoretical foundation and broad practical relevance, OT has been widely applied in diverse areas, including traffic control (Carlier et al., 2008; Danila et al., 2006; Barthélemy & Flammini, 2006), biomedical data analysis (Schiebinger et al., 2019; Koshizuka & Sato, 2022; Bunne et al., 2023), generative modeling (Wang et al., 2021; Onken et al., 2021; Zhang & Katsoulakis, 2023; Liu et al., 2019), and domain adaptation (Courty et al., 2016; 2017; Damodaran et al., 2018; Balaji et al., 2020). In recent years, there has been growing interest in deep learning techniques to solve OT problems, leading to the development of methods grounded in various mathematical formulations. Early approaches were primarily built upon the classical Monge formulation (Lu et al., 2020; Xie et al., 2019) and its relaxation into the Kantorovich framework (Makkuva et al., 2020). While theoretically rigorous, these methods often suffer from high computational complexity. The primal–dual formulation, which recasts the OT problem as a saddle-point optimization over the generative map and the Kantorovich potential function, has inspired scalable algorithms (Liu et al., 2019; Taghvaei & Jalali, 2019; Korotin et al., 2021a; Liu et al., 2021; Choi et al., 2024). Similar approaches have also been proposed for the Monge problem with general costs (Asadulaev et al., 2024; Fan et al., 2023). However, these approaches typically rely on adversarial training of two neural networks, which is challenging to manage and often introduces instability and inefficiency into the optimization process. Alternative approaches have investigated dynamical formulations using ordinary differential equations (ODEs) (Yang & Karniadakis, 2020; Onken et al., 2021; Tong et al., 2020; Huguet et al., 2022) and entropic-regularized models involving stochastic differential equations (SDEs) (Genevay et al., 2016; Seguy et al., 2017; Daniels et al., 2021; Gushchin et al., 2023; Zhou et al., 2024). Machine learning algorithms that unify Lagrangian and Eulerian perspectives of Mean Field Control problems Ruthotto et al. (2020); Lin et al. (2021); Zhao et al. (2025) likewise provide a computational framework for OT. Nevertheless, these methods typically require solving systems of differential equations, resulting in substantial computational overhead during both training and inference. Moreover, many existing methods yield bias maps that deviate from the OT solution due to the incorporation of regularization terms into the formulation.

Contributions. We propose a novel and efficient framework, termed *neural characteristic flow (NCF)*, for solving OT problems via the Hamilton–Jacobi (HJ) equation, whose viscosity solution characterizes the OT map. Despite its strong theoretical foundation for OT, the HJ formulation poses two major challenges: non-uniqueness of solutions and the need to solve ODEs in dynamical

Method (representative reference)	**Optimization**	**# Networks**	**OT direction**	**Sampling**	**Optimality of T**
Dual Formulation (Asadulaev et al., 2024)	Min-Max	Two	One-way	Direct	No
Dynamical Models (Onken et al., 2021)	Min	Single	Bidirectional	Iterative	No
HJ-based (Proposed)	Min	Single	Bidirectional	Direct	Yes

Table 1: Comparison of key features across different OT model approaches.

formulations. We overcome both by leveraging the method of characteristics and an implicit solution formula (Park & Osher, 2025) to obtain closed-form, bidirectional transport maps without numerical integration of ODEs. NCF uses a single neural network and avoids adversarial training or dual-network architectures, reducing complexity while improving efficiency. Our framework guarantees theoretical consistency with OT optimality conditions and supports a broad class of cost functions, including class-conditional transport. We also provide convergence analysis for Gaussian settings and demonstrate strong empirical performance across datasets of varying dimensions. A comparison of key features across different OT model approaches is summarized in Table 1.

2 Preliminary

2.1 Monge's Optimal Transport Problem

For a domain $\Omega \subset \mathbb{R}^d$, we denote $\mathscr{P}(\Omega)$ as the space of probability measures on Ω. Let $c : \Omega \times \Omega \to [0, \infty]$ be a cost function that measures the cost of transporting one unit of mass. For $\mu, \nu \in \mathscr{P}(\Omega)$, the classical Monge problem formulates OT as finding a measurable map $T : \Omega \to \Omega$ that pushes forward μ to ν, i.e., $T_\sharp \mu = \nu$, while minimizing the transportation cost:

$$W_c(\mu, \nu) \coloneqq \inf_{T_\sharp \mu = \nu} \int_\Omega c(\mathbf{x}, T(\mathbf{x}))\, \mathrm{d}\mu(\mathbf{x}). \tag{1}$$

We call a solution T^* to (1) an OT map between μ and ν. In the case where the cost c is expressed as a function of the difference between the two variables, T^* is characterized as follows:

Theorem 2.1 (Santambrogio (2015))**.** *When $c(\mathbf{x}, \mathbf{y}) = \ell(\mathbf{x} - \mathbf{y})$ for a lower semi-continuous (l.s.c.), sub-differentiable, and strictly convex function $\ell : \Omega \to \mathbb{R}$, the optimal map is expressed in terms of the Kantorovich dual potential function $\varphi^* : \Omega \to \mathbb{R}$ as*

$$T^*(\mathbf{x}) = \mathbf{x} + \nabla h(\nabla \varphi^*(\mathbf{x})), \tag{2}$$

where $h(\mathbf{z}) = \sup_{\mathbf{y} \in \mathbb{R}^d} \{\mathbf{z}^\top \mathbf{y} - \ell(\mathbf{y})\}$ is the Legendre transform of ℓ.

2.2 Dynamical Formulation

Benamou & Brenier (2000) formulate the OT (1) in a continuous-time dynamical formulation:

$$\inf_v \mathbb{E}_\mu \left[\int_0^{t_f} \ell(v(\mathbf{x}(t), t))\, \mathrm{d}t \right] \tag{3}$$

$$\text{s.t. } \dot{\mathbf{x}} = v,\ \mathbf{x}(0) \sim \mu,\ \mathbf{x}(t_f) \sim \nu, \tag{4}$$

where the terminal time $t_f > 0$ is typically set to 1. Within this dynamical framework, the associated optimality condition is governed by the *Hamilton–Jacobi (HJ) equation*:

$$\begin{cases} \frac{\partial u}{\partial t} + h(\nabla u) = 0 & \text{in } \Omega \times (0, t_f) \\ u = g & \text{on } \Omega \times \{t = 0\}, \end{cases} \tag{5}$$

coupled with the continuity equation that governs the evolution of the probability distribution. Here, ∇u denotes the gradient of u with respect to the spatial variable $\mathbf{x}$, and g represents the initial condition, whose explicit analytic form is typically intractable. The optimal velocity field is then determined by $v^* = \nabla h(\nabla u)$, where u is the *viscosity solution* to HJ equation (5).

3 Related Works

Deep learning methods for OT have gained traction following the development of scalable OT solvers (Genevay et al., 2016; Seguy et al., 2017) and WGANs (Arjovsky et al., 2017). Many approaches utilize GAN-based models to approximate OT plans, although they often suffer from training instability and extensive hyperparameter tuning. Another major line of work is based on the Kantorovich dual formulation (Kantorovich, 2006), where the OT map is recovered via optimization

of dual potentials, typically parameterized by input convex neural networks (ICNNs) (Amos et al., 2017). While theoretically sound, these methods involve unstable min-max optimization. To address these issues, natural gradient methods have been proposed to improve computational efficacy (Shen et al., 2020; Liu et al., 2024). Regularization techniques such as L^2 penalties (Genevay et al., 2016; Sanjabi et al., 2018) and cycle-consistency constraints (Korotin et al., 2019; 2021b) have been proposed, though unconstrained alternatives have shown stronger empirical performance (Korotin et al., 2021a; Fan et al., 2022).

To address the settings where deterministic OT maps may not exist, recent work has considered weak OT formulations (Backhoff-Veraguas et al., 2019). Neural approaches for weak OT and class-conditional transport have been proposed (Korotin et al., 2023; Asadulaev et al., 2024), but may yield spurious solutions under weak quadratic costs. Kernalized costs (Korotin et al., 2022) have been introduced to mitigate this.

OT has also been modeled as a dynamical system via continuous flows (Yang & Karniadakis, 2020; Tong et al., 2020; Onken et al., 2021; Huguet et al., 2022). While expressive, these methods require solving ODEs during training and inference, making them computationally expensive. Entropic and f-divergence regularized stochastic models (Daniels et al., 2021; Gushchin et al., 2023) improve smoothness but often rely on Langevin dynamics, which can be biased in high dimensions (Korotin et al., 2019). The HJ equation has been used to improve OT models, with physics-informed neural network (PINN) (Raissi et al., 2019) approaches applying L^2 penalties on HJ residuals to improve continuous normalizing flows, ODE-based formulations (Yang & Karniadakis, 2020; Onken et al., 2021), and stochastic variants (Zhang & Katsoulakis, 2023). However, due to the ill-posed nature of the HJ equation, this approach lacks guarantees for recovering the viscosity solution.

4 HJ Characteristic Flows for OT

In this section, we represent the OT map through the characteristics of the HJ equation, offering a principled and efficient framework for OT. Note that solving the HJ equation directly is challenging due to its inherent ill-posedness, non-smoothness of solutions, and gradient discontinuities, all of which complicate both theoretical analysis and numerical approximation.

Method of Characteristics. The viscosity solution to (5) is theoretically characterized by the following system of *characteristic ordinary differential equations (CODEs)*:

$$\begin{cases} \dot{\mathbf{x}} = \nabla h\left(\mathbf{p}\right) & \text{(6a)} \\ \dot{u} = -h(\mathbf{p}) + \mathbf{p}^\top \nabla h(\mathbf{p}) & \text{(6b)} \\ \dot{\mathbf{p}} = 0, & \text{(6c)} \end{cases}$$

where $\mathbf{p}$ denotes the shorthand for ∇u. CODE for $\mathbf{p}$ (6c) implies that $\mathbf{p}$ remains constant along each characteristic trajectory. Consequently, the characteristics are straight lines of the form $\mathbf{x}(t) = t\nabla h(\mathbf{p}) + \mathbf{x}(0)$, which coincide with the OT map in (2) at terminal time $t = t_f$. From a dynamical perspective, the ODE (4) can be interpreted as the characteristic equations (6a) of the HJ equation that determine the OT map (2). In other words, the transported point $T^*\left(\mathbf{x}\right)$ of a sample $\mathbf{x} \sim \mu$ corresponds to the terminal position of the characteristic line that originates from $\mathbf{x}$.

Our CODE formulation not only provides a principled construction of the forward transport map but also naturally characterizes the backward map. We denote by $T_\mu^{\nu *}$ the forward OT map transporting μ to ν, and by $T_\nu^{\mu *}$ the backward map transporting ν to μ.

Proposition 4.1 (Bidirectional OT Map)**.** *There exists a viscosity solution u^* to the HJ equation* (5) *that characterizes both the forward and backward OT maps through its forward and backward characteristic flows:*

$$T_\mu^{\nu *}(\mathbf{x}) = \mathbf{x} + t_f \nabla h\left(\nabla u^*\left(\mathbf{x}, 0\right)\right), \quad \mathbf{x} \sim \mu, \tag{7}$$

$$T_\nu^{\mu *}\left(\mathbf{y}\right) = \mathbf{y} - t_f \nabla h\left(\nabla u^*\left(\mathbf{y}, t_f\right)\right), \quad \mathbf{y} \sim \nu. \tag{8}$$

Accordingly, the viscosity solution of the HJ equation enables a bidirectional characterization of the OT map via forward and backward characteristic flows. Notably, since the characteristics are straight lines, both the forward and inverse transport maps admit explicit closed-form expressions. This obviates the need for numerical integration of ODEs typically required in conventional dynamical formulations. Consequently, the CODE-based formulation addresses a key computational bottleneck, enabling efficient and direct computation of bidirectional transport maps.

Implicit Solution Formula. Recently, a novel mathematical formulation for the viscosity solution of HJ equations has been developed using the system of CODEs (Park & Osher, 2025). Within this formulation, the viscosity solution admits the following implicit formula:

$$u(\mathbf{x}, t) = -th(\nabla u) + t\nabla u^\top \nabla h(\nabla u) + g(\mathbf{x} - t\nabla h(\nabla u)). \tag{9}$$

Proposition 4.2. *For OT problems* (1) *where ℓ satisfies the conditions in Theorem 2.1, the implicit solution formula* (9) *characterizes the viscosity solution of the HJ equation* (5) *almost everywhere.*

Proof. Detailed proof is provided in Appendix A.1. □

5 METHODS

5.1 OT WITH GENERAL COSTS

We propose a novel deep learning method, termed *neural characteristic flow (NCF)*, for learning bidirectional OT maps under general cost ℓ by solving the HJ equation (5) vis its implicit solution formula (9). The HJ equation characterizes the OT map as the gradient of the viscosity solution, ensuring that the resulting map minimizes the given cost functional. When coupled with the continuity equation, it also describes the evolution of probability distributions, thus guaranteeing correct mass transport from source to target. However, jointly solving this coupled system of PDEs is computationally expensive. To address this, the proposed NCF computes the OT map solely through the HJ equation, avoiding the need to solve the continuity equation explicitly.

Implicit Neural Representation. We represent the solution u of the HJ equation using an implicit neural representation (INR) $u_\theta : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ parameterized by θ. The network takes the spatial variable $\mathbf{x}$ and temporal variable t as input. By the universal approximation theorem (Hornik et al., 1989; Leshno et al., 1993), the INR can approximate the viscosity solution to the HJ equation. We denote by $T_\mu^\nu[u_\theta]$ as the transport map that aims to map μ to ν defined by (7) through u_θ:

$$T_\mu^\nu[u_\theta](\mathbf{x}) = \mathbf{x} + t_f \nabla h(\nabla u_\theta(\mathbf{x}, 0)). \tag{10}$$

The backward map $T_\nu^\mu[u_\theta]$ is analogously defined according to (8) via u_θ evaluated at $t = t_f$.

HJ-based Training Loss. While the HJ equation does not directly encode distributional information, it can recover the desired OT map, provided that an appropriate initial function g reflects the relationship between the source and target distributions. However, in practice, where only finite samples from these distributions are available, deriving an analytic form for g is generally intractable. To address this challenge, we introduce a loss term to ensure that the initial condition is appropriately learned during training, thereby steering the HJ solution toward accurately solving the desired OT problem. Specifically, this term enforces alignment between the generated samples obtained via $T[u_\theta]$ and the given target data. This alignment can be effectively quantified using discrepancy measures such as the maximum mean discrepancy (MMD) (Smola et al., 2006), whose value between two distributions μ and ν are defined as follows:

$$\mathrm{MMD}(\mu, \nu)^2 = \iint_{\Omega\times\Omega} k(\mathbf{x}, \mathbf{y})\, \mathrm{d}(\mu(\mathbf{x}) - \nu(\mathbf{x}))\, \mathrm{d}(\mu(\mathbf{y}) - \nu(\mathbf{y})), \tag{11}$$

where $k(\cdot,\cdot) : \Omega \times \Omega \to \mathbb{R}$ is a kernel function. The population loss for the MMD is

$$\mathcal{L}_{\mathrm{MMD}}(u_\theta) = \mathrm{MMD}(T_\mu^\nu[u_\theta]_\sharp \mu, \nu)^2. \tag{12}$$

We adopt the negative distance kernel $k(\mathbf{x}, \mathbf{y}) = -\|\mathbf{x} - \mathbf{y}\|_2$, which has proved to handle high-dimensional problems efficiently (Hertrich et al., 2024). With this kernel, the MMD loss becomes the squared energy distance (Rizzo & Székely, 2016).

In our implementation of the implicit solution formula, we replace the initial function g with u_θ evaluated at $t = 0$, and train the model using the following ϱ-weighted loss function

$$\mathcal{L}_{\mathrm{HJ}}(u_\theta) = \iint_{\Omega\times[0,t_f]} \Big(u_\theta + th(\nabla u_\theta) - t\nabla u_\theta^\top \nabla h(\nabla u_\theta) - u_\theta(\mathbf{x} - t\nabla h(\nabla u_\theta), 0)\Big)^2 \mathrm{d}\varrho(\mathbf{x})\, \mathrm{d}t, \tag{13}$$

where ϱ denotes a probability distribution on Ω.

The overall loss combines the implicit HJ loss and the MMD loss with a weight $\lambda > 0$:

$$\mathcal{L}_{\mathrm{HJ}}(u_\theta) + \lambda \mathcal{L}_{\mathrm{MMD}}(u_\theta). \tag{14}$$

We refer to Appendix B for practical choices of ϱ and the Monte Carlo estimation of the loss.

Advantages of the Proposed Approach. Our method offers several key advantages over existing OT frameworks, as summarized in Table 1. First, it jointly learns both forward and backward OT maps using a single neural network in one training phase. This contrasts with prior methods that require multiple networks, either due to the lack of invertibility or the use of adversarial dual formulations—leading to increased model complexity and training cost. Our method also avoids the instability of min-max optimization common in dual approaches, resulting in more stable training. Second, unlike dynamical OT models that require solving ODEs or SDEs, we use the method of characteristics to obtain OT maps in closed form. This removes the need for iterative solvers and improves sampling efficiency at both training and inference time. Third, our model directly incorporates the HJ equation via an implicit solution formula that reliably recovers the viscosity solution, as supported by the numerical results in Section 6. This not only aligns with the theoretical optimality conditions of OT but also helps identify and correct deviations from the target solution during training. Finally, our framework supports a broad class of cost functions beyond the quadratic case, offering greater flexibility and wider applicability across OT tasks.

5.2 Theoretical Analyses

In this section, we present theoretical analyses of our method, focusing on the OT problem with $\Omega = \mathbb{R}^d$ and the quadratic cost $\ell(\cdot) = \frac{1}{2}\|\cdot\|^2$, for which the corresponding Hamiltonian is given by $h(\cdot) = \frac{1}{2}\|\cdot\|^2$ as well. We prove that the minimizer of the loss (14) exactly recovers the true OT maps. Moreover, in the Gaussian setting, we establish stability analysis by showing that a small loss guarantees convergence to the true solution.

Consistency Analysis With some mild convexity assumption, we establish that the minimizer of (14) leads precisely to the optimal transport map.

Theorem 5.1 (Consistency of loss)**.** *Suppose the probability distributions μ, ν have finite second moments and $\varrho \in \mathscr{P}(\mathbb{R}^d)$ is strictly positive. Assume $u \in C^1_{loc}(\mathbb{R}^d \times [0, t_f])$, and define $u_1(\cdot) := u(\cdot, t_f) \in C^2_{loc}(\mathbb{R}^d)$ with $\nabla u_1 \in L^2(\mathbb{R}^d, \mathbb{R}^d; \nu)$. If u minimizes the loss functional* (14)*, i.e.,*

$$\mathcal{L}_{HJ}(u) + \lambda \mathcal{L}_{MMD}(u) = 0,$$

and the map $T^\mu_\nu[u]$ is bijective with its Jacobian $D_x T^\mu_\nu[u](x)$ is positive definite for any $x \in \mathbb{R}^d$, then $T^\mu_\nu[u]$ and $T^\nu_\mu[u]$ are the optimal transport maps from ν to μ, and vice versa.

The proof is provided in Appendix A.2. See also Remark A.5 for further discussion on the monotonicity condition for $D_x T^\mu_\nu[u]$.

Remark 5.2 (On regularity assumption of u). It is worth noting that the transport curves associated with the Wasserstein-2 OT problem do not intersect for $t \in [0, t_f]$ (cf. Chap. 8 of (Villani et al., 2008)). Since these curves constitute the characteristics of the HJ equation associated with the OT problem, we can expect classical solutions to the HJ equation, provided that μ and ν admit sufficiently regular density functions. This observation motivates the regularity assumption on u in Theorem 5.1. Moreover, u is parametrized with neural networks in practice, which naturally preserve the regularity.

Stability Analysis The loss (14) also exhibits favorable stability properties, which we illustrate in the Gaussian setting. Let $\mu = N(\mathbf{b}_\mu, \Sigma_\mu)$, $\nu = N(\mathbf{b}_\nu, \Sigma_\nu)$, then the OT map is

$$T^{\nu*}_\mu(\mathbf{x}) = A(\mathbf{x} - \mathbf{b}_\mu) + \mathbf{b}_\nu, \tag{15}$$

where $A := \Sigma_\mu^{-\frac{1}{2}}(\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_\mu^{-\frac{1}{2}}$. For analytical tractability, we consider a simplified quadratic parameterization $u_\theta(\mathbf{x}, t) = -(\frac{1}{2}\mathbf{x}^\top\theta_2(t)\mathbf{x} + \theta_1(t)^\top\mathbf{x} + \theta_0(t))$, where $\theta = [\theta_2(\cdot), \theta_1(\cdot), \theta_0(\cdot)] : [0, t_f] \to \mathbb{R}^{d\times d}_{\text{sym}} \times \mathbb{R}^d \times \mathbb{R}$. Although this represents a restricted subclass of neural networks, it permits rigorous analysis and yields insights relevant to more general architectures.

Assumption 5.3. $\theta(t)$ is bounded by K and K-Lipschitz. $\|\mathbf{b}_\mu\|, \|\mathbf{b}_\nu\|, \|\Sigma_\mu\|_F, \|\Sigma_\nu\|_F \le K$. A is strictly positive definite with smallest eigenvalue $\lambda_A > 0$.

Theorem 5.4 (Stability of loss)**.** *Under Assumption 5.3, the errors for u_θ and $T^\nu_\mu[u_\theta]$ satisfy*

$$\|u_\theta - u^*\|_{L^\infty([-1,1]^d)} + \left\|T^\nu_\mu[u_\theta] - T^{\nu*}_\mu\right\|_{L^\infty([-1,1]^d)} \le C\left(\mathcal{L}^{\frac{1}{3}}_{HJ} + \mathcal{L}^{\frac{1}{4}}_{MMD}\right), \tag{16}$$

where u^* and $T^{\nu*}_\mu$ are the true solution and OT map. C only depends on d, K and λ_A.

The theorem implies that sufficiently small loss guarantees convergence of the approximate solution u_θ—and consequently the resulting transport map $T_\mu^\nu[u_\theta]$—to their true counterparts. Furthermore, the proof shows that while multiple transport maps may minimize the MMD loss, the implicit HJ loss ensures that the OT map is uniquely recovered. The detailed description and proof for the theorem are deferred to Appendix A.3.

5.3 Class-Conditional OT

We extend our HJ-based framework to class-conditional OT, transporting source to target independently within each of the K labeled classes so as to preserve label consistency and class-specific structure. This formulation is particularly well-suited for domain adaptation and class-conditional generative modeling, where preserving class-specific features is crucial.

The OT map between samples of the k-th class must satisfy the HJ equation within the support of the corresponding class-specific distribution, as dictated by the optimality condition. Consequently, the global transport map $T_\mu^{\nu*}$ satisfies the HJ equation (5) across the entire domain. Although non-differentiable regions may arise due to intersections between transport maps of different classes, such discontinuities occur primarily in the boundaries between class supports. Since the gradient of the HJ solution is computed only within the support of each class-specific distribution, the transport map remains expressible in these regions. Accordingly, we retain the implicit HJ loss function (13) and modify the MMD loss to account for class conditioning as follows:

$$\mathcal{E}_{\text{class}}(\left(T_\mu^\nu\left[u_\theta\right]\right)_\sharp \mu, \nu) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{E}(\left(T_\mu^\nu\left[u_\theta\right]\right)_\sharp \mu_k, \nu_k). \tag{17}$$

A similar approach was proposed by Asadulaev et al. (2024).

6 Experimental Results

We evaluate the effectiveness of the proposed *neural characteristic flow (NCF)* across diverse OT tasks. All experiments in this section employ the quadratic cost function $\ell = \frac{1}{2}\|\cdot\|_2^2$, which is the canonical cost associated with the Wasserstein-2 distance. Computations were performed on a single NVIDIA GV100 (TITAN V) GPU. Further implementation details are provided in Appendix B.

6.1 Unconditional OT

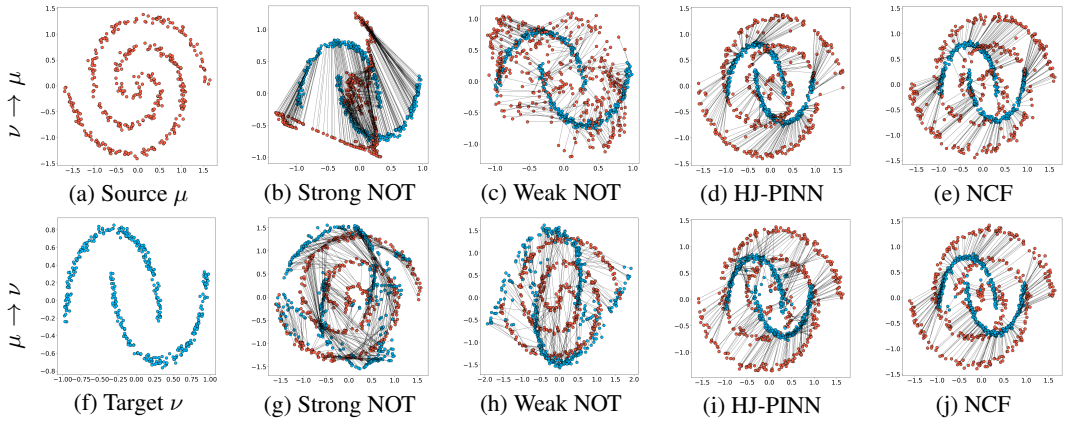
6.1.1 2D Toy Examples

We test the proposed NCF on a 2D toy dataset. We also compare our model with the neural optimal transport (NOT) framework (Korotin et al., 2023), including both the strong (deterministic) and weak (stochastic) variants. Since NOT directly parameterizes the transport map, it requires separate training for each transport direction. Additionally, we include an ablation study replacing our implicit solution formula loss (13) with a PINN loss on the HJ equation, referred to as HJ-PINN.

Figure 1 shows bidirectional transport results on 2D distributions. In addition to visualizing the transported distributions, we overlay the learned transport maps as black solid lines to assess whether each model has captured an OT plan. For weak NOT, the map is the average over noise inputs, as in the original work. Compared to all baselines, our method captures source and target distributions more accurately and learns transport maps closely aligned with the optimal solution. Strong NOT produces noisy, incoherent transport. Weak NOT performs better but still shows overlapping trajectories, indicating an incomplete OT representation. HJ-PINN yields noisy, intersecting transport paths, suggesting failure to learn OT dynamics. In contrast, our model learns accurate OT maps without trajectory crossings. Moreover, unlike NOT, which requires four separate networks for bidirectional training, our method achieves more accurate bidirectional transport with a single network. These results highlight the superior accuracy and efficiency of our approach. For further experimental results on the 2D example, please refer to Appendix C.1.

6.1.2 Evaluation on High-dimensional Gaussians

For general distributions, the ground truth OT solution is unknown, making quantitative evaluation challenging. To enable precise assessment, we consider the Gaussian case: $\mu = \mathcal{N}(\mathbf{0}, \Sigma_\mu)$ and $\nu = \mathcal{N}(\mathbf{0}, \Sigma_\nu)$, where a closed-form solution is available via (15). Following Korotin et al. (2021a), we vary the dimension d from 2 to 64, with Σ_μ and Σ_ν generated using random eigenvectors uniformly sampled on the unit sphere and logarithms of eigenvalues drawn uniformly from $[-2, 2]$.



(a) Source μ (b) Strong NOT (c) Weak NOT (d) HJ-PINN (e) NCF

(f) Target ν (g) Strong NOT (h) Weak NOT (i) HJ-PINN (j) NCF

Figure 1: *Swiss roll* (μ) $\rightleftarrows$ *Double moons* (ν): The top row shows transport in the direction $\nu \to \mu$, and the bottom row shows $\mu \to \nu$. The leftmost column displays μ and ν for reference.

Table 2: *Quantitative evaluation on Gaussian distributions.* UVP ($\downarrow$) is measured across different OT methods as the data dimension d increases.

Method	$d = 2$	$d = 4$	$d = 8$	$d = 16$	$d = 32$	$d = 64$
NOT	77.248	125.419	114.056	176.086	182.287	196.831
WGAN-QC	1.596	5.897	31.0367	59.314	113.237	141.407
LS	5.806	9.781	15.963	25.232	41.445	55.360
MM-v1	0.161	0.172	0.173	0.210	**0.374**	**0.415**
HJ-PINN	0.080	0.069	0.163	0.458	0.576	1.683
NCF	**0.010**	**0.021**	**0.086**	**0.146**	0.436	0.858

In addition to strong NOT and HJ-PINN, we evaluate several established OT methods: MM-v1 (Taghvaei & Jalali, 2019; Korotin et al., 2021a), which solves a min-max dual problem using input-convex neural networks (ICNNs); LS (Seguy et al., 2017), which addresses the dual problem via entropic regularization; and WGAN-QC (Liu et al., 2019), which employs a WGAN architecture with quadratic cost. Except for NOT—which directly parameterizes transport maps—all models use a shared architecture for potential functions.

Performance is measured using the unexplained variance percentage (UVP) (Korotin et al., 2019), which quantifies the L^2 error of the estimated transport map, normalized by $\mathrm{Var}(\nu)$. Computational efficiency is also evaluated in terms of training and inference time, peak memory usage, and memory required to store bidirectional OT maps. Table 2 reports UVP across models and dimensions, while Figure 2 summarizes computational metrics. Our method consistently yields accurate OT maps with favorable scaling behavior, outperforming NOT, WGAN-QC, and LS, which exhibit greater deviation from the ground-truth transport. While MM-v1 achieves marginally lower UVP in higher dimensions, it incurs over 20× longer training time and significantly higher memory usage. In contrast, our approach avoids expensive nested min-max optimization and leverages a single network, resulting in faster and more memory-efficient training. At inference, NOT offers the lowest latency due to its direct map parameterization, whereas other methods, including ours, require gradient-based evaluation, introducing additional overhead. This overhead, however, decreases with increasing dimension. Lastly, comparison with HJ-PINN underscores the superior effectiveness of our implicit loss in approximating the viscosity solution to the underlying HJ equation.

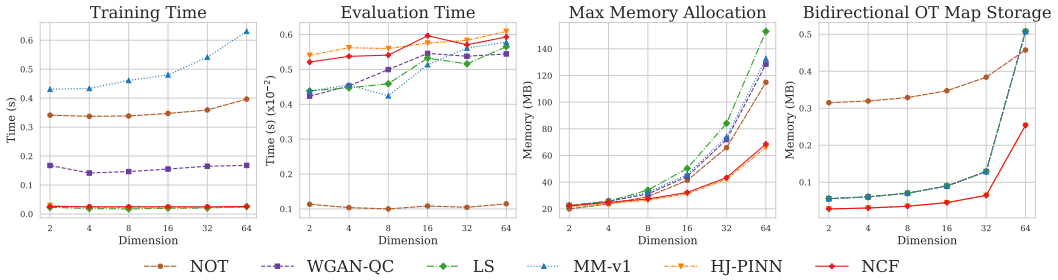


Figure 2: *Computational comparison.* Training time (s/epoch), evaluation time (s/epoch), peak memory (MB) during training, and memory (MB) for storing bidirectional OT maps are reported.

Table 3: *Quantitative evaluation of color transfer.* Earth mover distance (EMD) and histogram intersection (HI) between color distributions of target and transported images are reported.

Method	Winter-Summer		Monet-Photograph		Gogh-Photograph	
	EMD (↓)	HI (↑)	EMD (↓)	HI (↑)	EMD (↓)	HI (↑)
HisMatching	0.0012	0.7296	0.0013	0.7532	0.0010	0.7668
Reinhard	0.0013	0.6255	0.0012	0.7255	0.0009	0.7406
NOT	0.0008	0.8002	0.0008	0.8210	0.0008	0.8247
MM-v1	0.0014	0.7295	0.0011	0.7722	0.0007	0.8265
NCF	**0.0005**	**0.8914**	**0.0004**	**0.9174**	**0.0003**	**0.9117**

6.1.3 Application to Color Transfer

We employ the dataset provided by CycleGAN (Zhu et al., 2017) for image color transfer experiments. From each of the three available groups of image pairs, we selected 10 representative pairs. For each pair, we perform both forward and backward color transfer. To evaluate the effectiveness of our model, we include comparisons with two widely used classical color transfer methods: a standard per-channel histogram matching technique and the approach of Reinhard et al. (2001), which aligns the mean and standard deviation of color channels. These baselines represent statistical methods that do not rely on OT, providing a complementary perspective on performance. We include NOT and MM-v1 as deep learning OT baselines.

To quantitatively evaluate color fidelity and distributional consistency, we employ two widely used histogram-based metrics: Earth Mover's distance (EMD) and histogram intersection (HI), summarized in Table 3. Across all three domains, our method consistently achieves superior performance compared to all baselines in both metrics. In particular, our proposed method exhibits superior robustness in handling more complex and multimodal color distributions compared to MM-v1, especially in contrast to the simpler Gaussian settings examined in the previous section. Qualitative results are provided in Appendix C.2.

6.2 Class-Conditional OT

6.2.1 2D Toy Examples

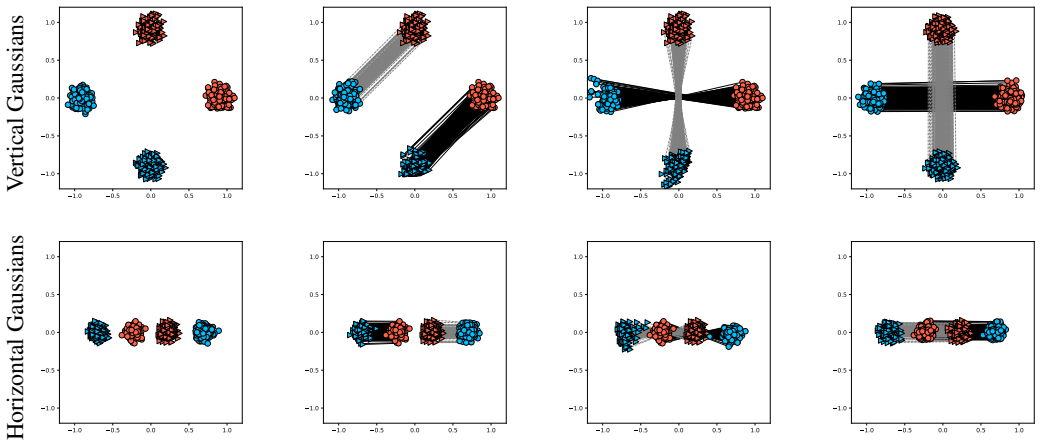
We present experimental results on a 2D synthetic dataset consisting of class-labeled samples, designed to evaluate class-conditional OT. To assess the ability of the proposed class-conditional NCF variant to model class-guided transport, we compare it against an unconditional NCF, which does not utilize label information. Furthermore, to benchmark our method against existing approaches, we include NOT with general cost functionals (GNOT) (Asadulaev et al., 2024), a recent model designed to perform class-conditional OT.

Figure 3 presents results on a 2D Gaussian mixture dataset, where each data point is associated with a class label. The unconditional NCF, lacking access to label information, learns a global transport map that ignores class structure, aligning source and target points purely based on W^2 distance. In contrast, both GNOT and the proposed class-conditional NCF learn separate transport maps per class. However, GNOT exhibits intersecting transport paths between classes, suggesting suboptimality with respect to the transport cost. The class-conditional NCF effectively disentangles transport across classes and yields maps that closely approximate the optimal solutions. These results highlight the accuracy and effectiveness of our approach, grounded in a CODE-based formulation of the HJ equation, for learning class-conditional transport in structured settings.

6.2.2 MNIST & Fashion MNIST

We apply our model to the MNIST (LeCun, 1998) and Fashion MNIST (Xiao et al., 2017) datasets, each comprising 10 classes. Given their substantially lower intrinsic dimensionality relative to the ambient space (Pope et al., 2021), we solve class-conditional OT problems in latent spaces obtained via β-VAEs (Higgins et al., 2017); see Appendix B.4 for details.

We consider transport from each Fashion MNIST class to its corresponding MNIST class; additional class-conditional OT tasks on MNIST are provided in Appendix C.3. We compare against baselines from Asadulaev et al. (2024), including NOT and GNOT, as well as a domain adaptation OT method (Courty et al., 2016; Flamary et al., 2021) that uses discrete OT with label-supervised regularization. Additionally, we evaluate unsupervised image translation methods AugCycleGAN (Almahairi et al., 2018) and MUNIT (Huang et al., 2018).



(a) Distributions (b) Unconditional NCF (c) GNOT (d) Class-conditional NCF

Figure 3: *2D class-conditional OT.* The leftmost column displays μ (red) and ν (blue), with class labels indicated by distinct markers. In the remaining columns, blue dots denote transported samples, while solid black and dotted gray lines represent the learned transport maps for each class.

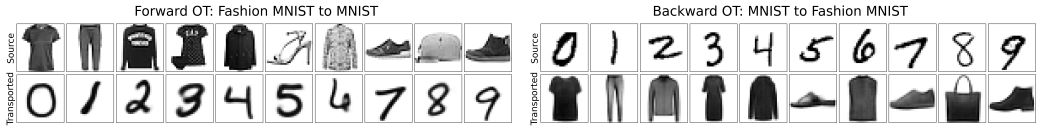


Figure 4: *Class-conditional OT between MNIST and Fashion MNIST*. **Left**: Forward OT. **Right**: Backward OT. The first row shows the source data, while the second row presents the data generated by learned OT map.

Table 4: Comparison of the accuracy and FID scores for the forward class-conditioned maps (Fashion MNIST $\rightarrow$ MNIST) learned using different methods. The accuracy and FID scores for the baseline methods are adopted from (Asadulaev et al., 2024).

Metric	**NOT** L^2 cost	**GNOT** Stochastic map	**Discrete OT** SinkhornLpL1	**AugCycleGAN**	**MUNIT**	**NCF [Ours]**
Accuracy(%) $\uparrow$	10.96	83.22	10.67	12.03	8.93	**83.42**
FID $\downarrow$	7.51	**5.26**	>100	26.35	7.91	18.27

Figure 4 shows bidirectional transported samples by NCF; uncurated results are in Appendix C.3. These results qualitatively demonstrate NCF's ability to perform bidirectional, class-conditional OT on real images. For quantitative evaluation, we report Fréchet Inception Distance (FID) (Heusel et al., 2017) and class-wise accuracy, which measures how well the class identity is preserved during transport, in Table 4. Our method achieves the highest accuracy, indicating its strong class-aware transport performance. Although the FID score is relatively high, this is largely due to the discrepancy introduced by the VAE decoder. To isolate this effect, we compute the FID between the NCF outputs and the VAE-decoded images. The resulting low score 2.73 indicates that the transport map in the latent space faithfully reproduces the target distribution. This is further supported by the KDE plots in Figure 14, showing close alignment between the transported and target latent distributions along principal components.

7 Conclusion

We introduced a theoretically grounded OT framework that recovers forward and backward maps in closed form via HJ characteristics. The resulting single-network, integration-free algorithm gives accurate, bidirectional maps, supports a broad class of costs, and extends to class-conditional transport with pairwise MMD alignment. We establish consistency and stability. Several tasks including synthetic, color-transfer, and MNIST demonstrate accuracy and efficiency of our algorithm.

Future directions include improving high-dimensional performance beyond latent-space implementations by developing more efficient gradient evaluations and scalable network designs. Extending the stability analysis to general neural architectures would provide a deeper theoretical understanding of our method and its convergence behavior.

REFERENCES

Amjad Almahairi, Sai Rajeshwar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *International conference on machine learning*, pp. 195–204. PMLR, 2018.

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International conference on machine learning*, pp. 146–155. PMLR, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Arip Asadulaev, Alexander Korotin, Vage Egiazarian, Petr Mokrov, and Evgeny Burnaev. Neural optimal transport with general cost functionals. *International conference on learning representations*, 2024.

Julio Backhoff-Veraguas, Mathias Beiglböck, and Gudmun Pammer. Existence, duality, and cyclical monotonicity for weak transport costs. *Calculus of Variations and Partial Differential Equations*, 58(6):203, 2019.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

Marc Barthélemy and Alessandro Flammini. Optimal traffic networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(07):L07002, 2006.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. *arXiv preprint arXiv:1807.07543*, 2018.

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768, 2023.

Guillaume Carlier, Cristian Jimenez, and Filippo Santambrogio. Optimal transportation with traffic congestion and wardrop equilibria. *SIAM Journal on Control and Optimization*, 47(3):1330–1350, 2008.

Jaemoo Choi, Yongxin Chen, and Jaewoong Choi. Improving neural optimal transport via displacement interpolation. *arXiv preprint arXiv:2410.03783*, 2024.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 447–463, 2018.

Max Daniels, Tyler Maunu, and Paul Hand. Score-based generative neural networks for large-scale optimal transport. *Advances in neural information processing systems*, 34:12955–12965, 2021.

Bogdan Danila, Yong Yu, John A Marsh, and Kevin E Bassler. Optimal transport on complex networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 74(4):046106, 2006.

Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational wasserstein gradient flow. *International Conference on Machine Learning*, 2022.

Jiaojiao Fan, Shu Liu, Shaojun Ma, Hao-Min Zhou, and Yongxin Chen. Neural monge map estimation and its applications. *Transactions on Machine Learning Research*, 2023.

Xue Feng and Thomas Strohmer. Improving autoencoder image interpolation via dynamic optimal transport. *arXiv preprint arXiv:2404.08900*, 2024.

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

Torres Fremlin. *Measure theory*, volume 4. Torres Fremlin, 2000.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

Nikita Gushchin, Alexander Kolesov, Alexander Korotin, Dmitry P Vetrov, and Evgeny Burnaev. Entropic neural optimal transport via diffusion processes. *Advances in Neural Information Processing Systems*, 36:75517–75544, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Johannes Hertrich, Christian Wald, Fabian Altekrüger, and Paul Hagemann. Generative sliced mmd flows with riesz kernels. *International conference on learning representations*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018.

Guillaume Huguet, Daniel Sumner Magruder, Alexander Tong, Oluwadamilola Fasina, Manik Kuchroo, Guy Wolf, and Smita Krishnaswamy. Manifold interpolating optimal-transport flows for trajectory inference. *Advances in neural information processing systems*, 35:29705–29718, 2022.

Leonid V Kantorovich. On the translocation of masses. *Journal of mathematical sciences*, 133(4), 2006.

Diederik Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. *arXiv preprint arXiv:1909.13082*, 2019.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *Advances in neural information processing systems*, 34:14593–14605, 2021a.

Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. *arXiv preprint arXiv:2102.01752*, 2021b.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Kernel neural optimal transport. *arXiv preprint arXiv:2205.15269*, 2022.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. *International conference on learning representations*, 2023.

Takeshi Koshizuka and Issei Sato. Neural lagrangian schr\" odinger bridge: Diffusion modeling for population dynamics. *arXiv preprint arXiv:2204.04853*, 2022.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

Alex Tong Lin, Samy Wu Fung, Wuchen Li, Levon Nurbekyan, and Stanley J Osher. Alternating the population and control neural networks to solve high-dimensional stochastic mean-field games. *Proceedings of the National Academy of Sciences*, 118(31):e2024713118, 2021.

Huidong Liu, Xianfeng Gu, and Dimitris Samaras. Wasserstein gan with quadratic transport cost. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4832–4841, 2019.

Shu Liu, Shaojun Ma, Yongxin Chen, Hongyuan Zha, and Haomin Zhou. Learning high dimensional wasserstein geodesics. *arXiv preprint arXiv:2102.02992*, 2021.

Shu Liu, Stanley Osher, and Wuchen Li. A natural primal-dual hybrid gradient method for adversarial neural network training on solving partial differential equations. *arXiv preprint arXiv:2411.06278*, 2024.

Guansong Lu, Zhiming Zhou, Jian Shen, Cheng Chen, Weinan Zhang, and Yong Yu. Large-scale optimal transport via adversarial training with cycle-consistency. *arXiv preprint arXiv:2003.06635*, 2020.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9223–9232, 2021.

Yesom Park and Stanley Osher. Neural implicit solution formula for efficiently solving hamilton-jacobi equations. 2025. URL https://arxiv.org/abs/2501.19351.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

Maria L Rizzo and Gábor J Székely. Energy distance. *wiley interdisciplinary reviews: Computational statistics*, 8(1):27–38, 2016.

Lars Ruthotto, Stanley J Osher, Wuchen Li, Levon Nurbekyan, and Samy Wu Fung. A machine learning framework for solving high-dimensional mean field game and mean field control problems. *Proceedings of the National Academy of Sciences*, 117(17):9183–9193, 2020.

Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. On the convergence and robustness of training gans with regularized optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

Zebang Shen, Zhenfu Wang, Alejandro Ribeiro, and Hamed Hassani. Sinkhorn natural gradient for generative models. *Advances in Neural Information Processing Systems*, 33:1646–1656, 2020.

Alexander J Smola, A Gretton, and K Borgwardt. Maximum mean discrepancy. In *13th international conference, ICONIP*, volume 6, 2006.

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

Amirhossein Taghvaei and Amin Jalali. 2-wasserstein approximation via restricted convex potentials with application to improved training for gans. *arXiv preprint arXiv:1902.07197*, 2019.

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pp. 9526–9536. PMLR, 2020.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, and Can Yang. Deep generative learning via schrödinger bridge. In *International conference on machine learning*, pp. 10794–10804. PMLR, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yujia Xie, Minshuo Chen, Haoming Jiang, Tuo Zhao, and Hongyuan Zha. On scalable and efficient computation of large scale optimal transport. In *International Conference on Machine Learning*, pp. 6882–6892. PMLR, 2019.

Liu Yang and George Em Karniadakis. Potential flow generator with l2 optimal transport regularity for generative models. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2): 528–538, 2020.

Benjamin J Zhang and Markos A Katsoulakis. A mean-field games laboratory for generative modeling. *arXiv preprint arXiv:2304.13534*, 2023.

Jiaxi Zhao, Mo Zhou, Xinzhe Zuo, and Wuchen Li. Variational conditional normalizing flows for computing second-order mean field control problems. *arXiv preprint arXiv:2503.19580*, 2025.

Mo Zhou, Stanley Osher, and Wuchen Li. Score-based neural ordinary differential equations for computing mean field control problems. *arXiv preprint arXiv:2409.16471*, 2024.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.