SIMULTANEOUS NEURAL OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimal Transport (OT) provides a principled framework for learning transformations between probability distributions from unpaired samples. In many applications, however, a single transformation must map several source distributions to a common target distribution. For example, image restoration might require handling different types of degradation without knowing the degradation of each input at inference time. Simple approaches of pooling the source distributions only encourage alignment with the target at the aggregate level and may leave individual sources misaligned. In our paper, we consider the simultaneous OT problem which formalizes the task of learning a shared transport map that minimizes the average transport cost while aligning each source distribution with a prescribed target. We propose a neural method for solving the simultaneous OT problem by learning a shared transport map that minimizes the average transport cost while aligning each source distribution with a prescribed target. We derive a max-min formulation for learning this map. We illustrate its application to image restoration, where a single model handles multiple degradation types using a common collection of clean target images.

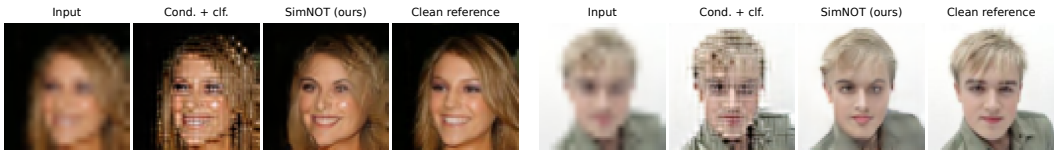

Figure 1: Restoration of selected CelebA 64×64 test images under bilinear downsampling with an unseen factor ($\times3$ at test time versus $\times4$ during training). Each group shows the degraded input, conditional OT with classifier-based routing, SimNOT (ours), and the clean reference.

1 INTRODUCTION

Learning transformations between probability distributions from unpaired samples underlies a range of applications, including image-to-image translation (Zhu et al., 2017), voice conversion (Chun et al., 2023), and prediction of single-cell perturbation responses (Bunne et al., 2023). In many practical scenarios, the input data originate from several distributions, while a single model is expected to transform them into a common target distribution. The challenge is to learn one transformation that produces suitable outputs for each source distribution while preserving relevant properties of the inputs. For example, all-in-one image restoration aims to handle multiple types of degradation within one model without requiring the degradation type at inference time (Potlapalli et al., 2023). In the unpaired setting, the learner has access to degraded and clean images but not to their correspondences.

Optimal Transport (OT) provides a mathematical framework for learning such transformations (Peyré et al., 2019). It seeks a transport map or plan that aligns the source and target distributions while minimizing a prescribed transport cost. The cost specifies which input–output correspondences are preferable and can be chosen to encourage preservation of particular attributes. Neural OT methods make it possible to approximate transport maps between distributions available only through samples and apply the learned maps to previously unseen inputs (Fan et al., 2023; Korotin et al., 2023a;b). These methods provide a natural starting point for learning transformations shared across several source distributions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A straightforward approach is to pool the source datasets and learn an OT map from their mixture to the target distribution. However, matching the transformed mixture to the target does not guarantee that the transformed distribution of each source matches the target individually. For example, different sources may be mapped to different parts of the target distribution, producing a correct aggregate distribution despite substantial discrepancies for individual sources. Alternatively, learning a separate map for each source does not yield a single transformation that can be applied without selecting a source-specific model. This motivates learning a shared map while explicitly accounting for each source distribution.

The *simultaneous optimal transport* (SOT) framework (Wang & Zhang, 2025) formalizes transport under multiple distributional constraints using a common map or stochastic kernel. We consider its many-to-one setting: a single transformation must align several source distributions with a prescribed target while minimizing the average transport cost. The shared transformation couples the source-specific transport tasks, and the individual alignment requirements retain information that is lost when the sources are treated only as a mixture. Our focus is on the computational setting in which the distributions are unknown and available through unpaired samples, and the learned transformation must generalize to new inputs.

To develop a practical learning method, we consider an *unbalanced* formulation of simultaneous transport, replacing hard marginal constraints with divergence penalties. Unbalanced OT has already been used to learn generative models and transport maps between a pair of distributions (Choi et al., 2023; Gazdieva et al., 2024; Yang & Uhler, 2018; Klein et al., 2024; Eyring et al., 2024). In our setting, this relaxation provides an adjustable trade-off between transport cost and marginal agreement while retaining the shared transport structure and separate objectives for the sources. We derive a variational training objective and use it to learn a single neural transport map together with source-specific potentials. The source identity is used to select the appropriate potential during training, whereas the transport map itself is shared and does not require this identity at inference time.

Contributions. We develop **Simultaneous Neural Optimal Transport (SimNOT)**, a continuous solver for learning a shared transport map from multiple source distributions to a prescribed target using unpaired samples. We introduce a divergence-based unbalanced formulation of simultaneous OT, establish its exact semi-dual representation (§4.1), and provide neural approximation guarantees (§4.3). These results yield a neural solver with one shared map and source-specific potentials, requiring no source labels at inference time (§4.2). We demonstrate its effectiveness on simultaneous Gaussian-to-Swiss-roll transport and unpaired image restoration with multiple degradation types, including evaluation under unseen degradation strengths (§5).

Notation. We take the source and target domains to be nonempty closed sets $\mathcal{X} \subseteq \mathbb{R}^{D_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{D_y}$. We write $\mathcal{Z}$ for an auxiliary noise space. For a Borel space E, $\mathcal{P}(E)$ and $\mathcal{M}_+(E)$ denote the sets of Borel probability measures and finite non-negative Borel measures on E, respectively, and $C_b(E)$ denotes the bounded continuous real-valued functions on E. The source distributions are denoted by $\mathbb{P}_1, \dots, \mathbb{P}_K \in \mathcal{P}(\mathcal{X})$, and the common target distribution by $\mathbb{P}^* \in \mathcal{P}(\mathcal{Y})$. For a measurable map T, $T_{\#}\mu$ denotes the pushforward of a measure μ. We represent a stochastic map by $T : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$, where $z \sim \mathbb{S} \in \mathcal{P}(\mathcal{Z})$ is independent of the input; its output distribution for source $\mathbb{P}_k$ is $T_{\#}(\mathbb{P}_k \otimes \mathbb{S})$. For a probability kernel $\gamma(\cdot \mid x)$ from $\mathcal{X}$ to $\mathcal{Y}$, we write $\gamma_{\#}\mu$ for the output measure $\int \gamma(\cdot \mid x)\, d\mu(x)$. For $\gamma \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$, its marginals are denoted by γ_x and γ_y. We use $\Pi(\mu, \nu)$ for the set of finite non-negative measures on the product space with marginals μ and ν (when these marginals have the same total mass); in particular, for probability marginals these are probability couplings.

2 Related Work

We review the most closely related OT formulations and solvers below. *Extended related work* on OT barycenters and all-in-one image restoration is discussed in Appendix B.

Learning optimal transport maps. OT has been used in generative modeling both as a discrepancy between generated and target distributions and as a principle for learning the transformation itself. For example, Wasserstein GANs (Arjovsky et al., 2017) use an OT discrepancy as a training loss, but do not require the generator to minimize the transport cost between its inputs and outputs. More closely related to our work are continuous neural OT solvers, which learn transport maps or plans from samples (Makkuva et al., 2020; Rout et al., 2022; Korotin et al., 2023b). These methods exploit OT

duality to construct optimization objectives for neural potentials and transport maps, with applications to generation and unpaired translation. Partial transport formulations also permit relaxing distribution matching while controlling the similarity between inputs and outputs (Gazdieva et al., 2023). These works consider transport between a pair of distributions; our focus is on learning one transformation jointly for several sources.

Simultaneous optimal transport. Wang & Zhang (2025) introduce simultaneous OT as transport between vector-valued measures: a common map or transport kernel must transport several source measures to their respective targets. They study Monge and Kantorovich formulations, existence, and duality. The balanced problem with a common target considered in our work is a particular case of this framework. Its defining requirement is that the same transformation satisfy the transport constraints for every source. Learning independent pairwise OT maps does not impose this requirement, while matching a pooled source distribution to the target does not ensure that each source is matched individually. We build on this perspective to develop a neural method for learning shared transformations from samples. We also consider divergence-based marginal relaxation; this differs from the inequality-constrained unbalanced formulation studied by Wang & Zhang (2025).

Unbalanced optimal transport. Unbalanced OT replaces exact marginal constraints with penalties for deviations from the prescribed measures (Liero et al., 2018). Among continuous solvers, Yang & Uhler (2018) propose an adversarial approach that jointly learns a transport map and a source mass scaling function. UOTM (Choi et al., 2023) learns a transport network and a potential through a semi-dual objective. The latter approach is particularly relevant to the optimization principles used in our work. Both methods consider transport between two measures. Our method combines marginal relaxation with a shared transformation and source-specific potentials. The role of the unbalanced formulation here is to provide a flexible transport objective for simultaneous learning.

3 Background

In this section, we recall the OT formulations relevant to our work and specify the simultaneous OT problem with a common target. We then introduce the marginal relaxation considered in this paper. For background on OT and unbalanced OT, we refer to Villani (2009) and Chizat (2017), respectively.

3.1 Optimal Transport

Let $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ be probability measures, $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ – continuous function.

Monge formulation. The Monge OT problem seeks a measurable map $T : \mathcal{X} \to \mathcal{Y}$ that transports $\mathbb{P}$ to $\mathbb{Q}$ while minimizing the transport cost given by function $c(\cdot, \cdot)$:

$$\inf_{T:\, T_{\#}\mathbb{P}=\mathbb{Q}} \int_{\mathcal{X}} c(x, T(x))\, d\mathbb{P}(x). \tag{1}$$

A minimizer T^* of equation 1, when it exists, is called an optimal transport (OT) map.

Kantorovich formulation. Monge OT problem may admit no feasible map or no minimizer. Thus, it is common to consider its Kantorovich relaxation which allows splitting the mass of each input point:

$$\mathrm{OT}_c(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, d\pi(x, y). \tag{2}$$

Here $\Pi(\mathbb{P}, \mathbb{Q})$ is the set of probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals $\pi_x = \mathbb{P}$ and $\pi_y = \mathbb{Q}$. Under the assumptions of this paper, a minimizer π^* of equation 2 always exists, but need not be unique (Villani, 2009). Such a minimizer is called an optimal transport plan.

Disintegration gives $d\pi(x, y) = d\pi_x(x)\, d\pi(y \mid x)$, where $\pi(\cdot \mid x)$ is the conditional distribution of outputs for an input x. An OT plan π^* thus defines a stochastic transport map $x \mapsto \pi^*(\cdot \mid x)$. If $\pi^*(\cdot \mid x) = \delta_{T^*(x)}$, then $\pi^* = (\mathrm{id}_{\mathcal{X}}, T^*)_{\#}\mathbb{P}$ and T^* is an OT map.

Unbalanced formulation. Unbalanced OT (UOT) relaxes the marginal constraints in equation 2 by penalizing deviations of the marginals from $\mathbb{P}$ and $\mathbb{Q}$. For finite non-negative measures μ, ν on the

same space, we define the ψ-divergence as

$$D_\psi(\mu\|\nu) = \begin{cases} \int \psi\left(\frac{d\mu}{d\nu}\right) d\nu, & \mu \ll \nu, \\ +\infty, & \text{otherwise.} \end{cases} \tag{3}$$

Here $\psi : [0,\infty) \to [0,\infty]$ is convex, lower semicontinuous, vanishes only at 1, and satisfies $\lim_{u\to\infty} \psi(u)/u = +\infty$. We define D_ϕ analogously. Examples include the Kullback–Leibler and χ^2 divergences, generated by $\psi(u) = u\log u - u + 1$ and $\psi(u) = (u-1)^2$, respectively.

The UOT problem is given by (Liero et al., 2018)

$$\mathrm{UOT}_{c,\psi,\phi}(\mathbb{P},\mathbb{Q}) = \inf_{\gamma\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \left[\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\gamma(x,y) + D_\psi(\gamma_x\|\mathbb{P}) + D_\phi(\gamma_y\|\mathbb{Q}) \right], \tag{4}$$

where $\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})$ is the set of finite non-negative measures, and γ_x, γ_y are the marginals of γ. A minimizer γ^* is called an unbalanced OT plan. Its total mass is not constrained to equal one. The strength of the marginal penalties can be adjusted by scaling the generators ψ and ϕ.

The balanced problem equation 2 is recovered when ψ and ϕ are the convex indicators of $\{1\}$: the divergence terms then enforce $\gamma_x = \mathbb{P}$ and $\gamma_y = \mathbb{Q}$.

As in the balanced case, disintegration gives $d\gamma(x,y) = d\gamma_x(x)\, d\gamma(y \mid x)$. The conditional distributions $\gamma^*(\cdot \mid x)$ define a stochastic transport map acting on the source marginal γ_x^*, which may differ from $\mathbb{P}$. In the deterministic case, $\gamma^*(\cdot \mid x) = \delta_{T^*(x)}$ and $\gamma_y^* = T^*_\# \gamma_x^*$.

3.2 Simultaneous Optimal Transport

Simultaneous OT seeks a shared transport rule that transports several source measures to their respective target measures (Wang & Zhang, 2025). Here, we consider the case of K source distributions $\mathbb{P}_1, \ldots, \mathbb{P}_K \in \mathcal{P}(\mathcal{X})$ and a common prescribed target distribution $\mathbb{P}^* \in \mathcal{P}(\mathcal{Y})$.

Balanced formulation. A simultaneous transport plan from $\mathbb{P}_1, \ldots, \mathbb{P}_K$ to $\mathbb{P}^*$ is represented by a family of plans $\pi_k \in \Pi(\mathbb{P}_k, \mathbb{P}^*)$ sharing a common conditional distribution $\pi(\cdot \mid x)$:

$$d\pi_k(x,y) = d\mathbb{P}_k(x)\, d\pi(y \mid x),\ k = 1, \ldots, K. \tag{5}$$

Thus, the same conditional distribution of outputs is used for a given input x across all sources. The balanced simultaneous OT problem seeks such a family minimizing the average transport cost:

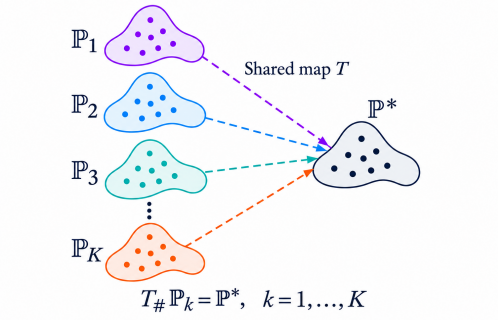

Figure 2: A schematic illustration of simultaneous OT.

$$\inf_{\substack{\pi(\cdot|x):\\ (\pi_k)_y = \mathbb{P}^*,\ k=1,\ldots,K}} \frac{1}{K} \sum_{k=1}^{K} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\pi_k(x,y). \tag{6}$$

Here, the infimum is taken over shared conditional probability distributions, with the plans π_k defined by equation 5. A minimizing family $\{\pi_k^*\}_{k=1}^K$, when it exists, is called a simultaneous optimal transport (OT) plan. Its shared conditional distribution $\pi^*(\cdot \mid x)$ defines a stochastic transport map $x \mapsto \pi^*(\cdot \mid x)$ that transports every $\mathbb{P}_k$ to $\mathbb{P}^*$. In the special case of deterministic plans, $\pi(\cdot \mid x) = \delta_{T(x)}$ for a shared map $T : \mathcal{X} \to \mathcal{Y}$ satisfying $T_\#\mathbb{P}_k = \mathbb{P}^*$ for all k, see Fig. 2. Optimizing equation 6 over such plans yields the simultaneous Monge problem.

When the source distributions have pairwise disjoint supports, simultaneous OT reduces to independent pairwise OT problems: the corresponding transport maps, when they exist, can be combined into a single map. For overlapping sources, independently obtained maps may disagree on the overlap, whereas simultaneous OT requires the same transformation for every source.

Unbalanced formulation. We consider an unbalanced relaxation of equation 6 that penalizes deviations from the prescribed marginals while retaining a shared conditional distribution. Specifically, we consider finite nonnegative transport plans $\gamma_1, \ldots, \gamma_K$ of the form

$$d\gamma_k(x,y) = d(\gamma_k)_x(x)\, d\gamma(y \mid x), \qquad k = 1, \ldots, K, \tag{7}$$

where $\gamma(\cdot \mid x)$ is a shared conditional probability distribution. The source marginals $(\gamma_k)_x$ are optimized jointly with this conditional distribution, and the plans γ_k need not have unit mass.

The resulting simultaneous UOT problem is

$$\inf_{\gamma(\cdot|x),\, \{(\gamma_k)_x\}_{k=1}^K} \frac{1}{K}\sum_{k=1}^{K}\left[\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, d\gamma_k(x,y) + D_\psi((\gamma_k)_x\|\mathbb{P}_k) + D_\phi((\gamma_k)_y\|\mathbb{P}^*)\right]. \quad (8)$$

Here, the infimum is taken over shared conditional probability distributions and finite nonnegative source marginals, with the plans γ_k defined by equation 7. The divergences penalize deviations of the source and target marginals from $\mathbb{P}_k$ and $\mathbb{P}^*$, respectively.

A minimizing family $\{\gamma_k^*\}_{k=1}^K$, when it exists, is called a simultaneous UOT plan. Its shared conditional distribution $\gamma^*(\cdot \mid x)$ defines the corresponding stochastic transport map. In the case of deterministic plans, $\gamma(\cdot \mid x) = \delta_{T(x)}$ for a shared map $T : \mathcal{X} \to \mathcal{Y}$, and $(\gamma_k)_y = T_\#(\gamma_k)_x$ for all k.

To the best of our knowledge, this divergence-based formulation of simultaneous unbalanced OT has not been studied previously. The unbalanced formulation of Wang & Zhang (2025) uses marginal domination constraints.

3.3 Computational Setup

The distributions $\mathbb{P}_1, \ldots, \mathbb{P}_K$ and $\mathbb{P}^*$ are unknown and accessible only through finite unpaired samples. Our goal is to learn a shared transport map for the underlying distributions that admits new inputs not present in the training data. This setup is typically called continuous (Fan et al., 2023; Rout et al., 2022; Korotin et al., 2023a) and differs from the discrete one, which seeks transport couplings between empirical measures (Cuturi, 2013; Chizat et al., 2018).

4 Simultaneous Neural OT Method

In this section, we derive our novel optimization objective for learning simultaneous UOT plans (§4.1) and propose an algorithm to solve (§4.2). We then study the approximation properties of shared neural maps (§4.3). The *proofs of all theoretical results* are given in Appendix A.

4.1 Deriving the Optimization Objective

Direct optimization of equation 8 is difficult from samples because the marginal divergences involve unknown measures. We therefore derive an equivalent semi-dual objective whose dependence on the data distributions is only through expectations.

For the result below, assume that $\mathbb{P}_1, \ldots, \mathbb{P}_K$ are atomless, $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ is finite and continuous, and that ψ and ϕ satisfy the assumptions stated in Section 3.1 and $\phi(0) < +\infty$, $\psi(0) < +\infty$. Put $\overline{\mathbb{P}} = \frac{1}{K}\sum_{k=1}^K \mathbb{P}_k$. This is equivalent to atomlessness of $\overline{\mathbb{P}}$: an average of atomless measures is atomless, while each $\mathbb{P}_k$ is absolutely continuous with respect to $\overline{\mathbb{P}}$. The atomlessness assumption is used only in the purification argument of Lemma 1. Let J^* denote the infimum in equation 8.

Theorem 1 (Semi-dual formulation of simultaneous UOT)**.** *Under the assumptions above,*

$$J^* = \sup_{\mathbf{v}\in C_b(\mathcal{Y})^K} \inf_{\gamma} \mathcal{J}(\mathbf{v}, \gamma) = \sup_{\mathbf{v}\in C_b(\mathcal{Y})^K} \inf_{T:\mathcal{X}\to\mathcal{Y}} \mathcal{J}(\mathbf{v}, T), \quad (9)$$

where the first infimum is over all probability kernels $\gamma(\cdot \mid x)$ from $\mathcal{X}$ to $\mathcal{Y}$ and

$$\mathcal{J}(\mathbf{v}, \gamma) = -\frac{1}{K}\sum_{k=1}^{K}\left[\mathbb{E}_{x\sim\mathbb{P}_k}\mathbb{E}_{y\sim\gamma(\cdot|x)}\bar{\psi}(v_k(y) - c(x,y)) + \mathbb{E}_{y\sim\mathbb{P}^*}\bar{\phi}(-v_k(y))\right]. \quad (10)$$

For a deterministic map T, we use the shorthand $\mathcal{J}(\mathbf{v}, T) := \mathcal{J}(\mathbf{v}, \delta_{T(\cdot)})$.

In equation 10, the distributions enter only through expectations and can be estimated from samples.

Remark on attainment. Theorem 1 concerns optimal values and does not require attainment of the primal infimum. Existence is assumed explicitly whenever it is needed. The deterministic representation in equation 9 refers only to the inner semi-dual problem; the structure of primal minimizers is addressed separately in Theorem 3.

4.2 Parametrization and Training

Parametrization. To optimize over the shared conditional distribution $\gamma(\cdot \mid x)$ in equation 9, we represent it using a stochastic or deterministic map, following Choi et al. (2023). Let $\mathcal{Z} \subseteq \mathbb{R}^{D_z}$ be a Borel set and $\mathbb{S} \in \mathcal{P}(\mathcal{Z})$ be an atomless noise distribution. Every shared conditional distribution $\gamma(\cdot \mid x)$ in equation 9 admits a measurable realization $T : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ with $z \sim \mathbb{S}$: $\gamma(\cdot \mid x)[= \gamma_T(\cdot \mid x)] = (T(x, \cdot))_{\#}\,\mathbb{S}$, where the map T and the noise distribution $\mathbb{S}$ are shared across all sources.

Using this representation, we rewrite the optimization problem equation 9 as

$$J^* = \sup_{v_1,\dots,v_K \in C(\mathcal{Y})} \inf_{T} \mathcal{J}(v_1, \dots, v_K, T), \tag{11}$$

where the infimum is over measurable maps $T : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ and

$$J(v_1, \dots, v_K, T) = -\frac{1}{K} \sum_{k=1}^{K} \Big[\mathbb{E}_{\substack{x \sim \mathbb{P}_k \\ z \sim \mathbb{S}}} \bar{\psi}(v_k(T(x, z)) - c(x, T(x, z))) + \mathbb{E}_{y \sim \mathbb{P}^*} \bar{\phi}(-v_k(y)) \Big]. \tag{12}$$

In some settings, stochasticity of the conditional distribution is not needed. We can then consider a shared measurable map $T : \mathcal{X} \to \mathcal{Y}$, which defines deterministic conditional distributions $\gamma_T(\cdot \mid x) = \delta_{T(x)}$. In this case, the expectation over $\mathbb{S}$ in equation 12 is omitted. Theorem 3 below motivates this choice when the transport cost is quadratic and both marginal penalties are KL divergences.

To solve equation 11, we parametrize the shared map and the potentials by neural nets $T_\theta : \mathcal{X} \times \mathcal{Z} \to Y$ and $v_{\omega_k} : \mathcal{Y} \to \mathbb{R}$, $k = 1, \dots, K$, with trainable parameters θ and $\omega_1, \dots, \omega_K$, respectively. The noise input is omitted for deterministic maps. In practice, we scale the transport cost c by an unbalancedness parameter $\tau > 0$: smaller values favor balanced simultaneous OT, while larger values allow greater marginal deviations.

Training. In our setup, the source distributions $\mathbb{P}_1, \dots, \mathbb{P}_K$ and the target distribution $\mathbb{P}^*$ are accessible only through samples. We therefore estimate the objective $\mathcal{J}(v_{\omega_1}, \dots, v_{\omega_K}, T_\theta)$ in equation 12 using Monte Carlo with random mini-batches from the source, target and noise distributions.

Training alternates between gradient ascent with respect to the potential parameters $\omega_1, \dots, \omega_K$ and gradient descent with respect to the map parameters θ. Each potential is trained using its corresponding source distribution and the common target, while the shared map receives contributions from all sources. The training procedure is summarized in Algorithm 1.

Inference. Given a new input x, we generate an output $T_\theta(x, z)$ with $z \sim \mathbb{S}$, or $T_\theta(x)$ in the deterministic case. The source index and the potentials are not required at inference time.

4.3 Theoretical Properties

The semi-dual formulation allows an arbitrary shared stochastic kernel. We first ask whether one neural map with a noise input is expressive enough to approximate such a kernel simultaneously for all source distributions.

For the approximation results in this subsection only, assume that $\mathcal{Y} = [a, b]^{D_y}$ for some $a < b$, and let $\mathbb{S} = \mathrm{Unif}[0, 1]$. The compact box assumption makes the W_2 statement automatic and allows coordinatewise clipping to $\mathcal{Y}$ to be implemented by ReLU operations.

Theorem 2 (Approximation by a shared neural map)**.** *Let $\gamma(\cdot \mid x)$ be a Borel probability kernel from $\mathcal{X}$ to $\mathcal{Y}$. For every $\varepsilon > 0$, there exists a finite fully connected ReLU network*

$$T_\theta : \mathbb{R}^{D_x + 1} \to \mathcal{Y}$$

such that, with

$$\gamma_\theta(\cdot \mid x) = (T_\theta(x, \cdot))_{\#}\,\mathbb{S},$$

we have simultaneously for every $k = 1, \dots, K$,

$$\int_{\mathcal{X}} W_2^2(\gamma_\theta(\cdot \mid x), \gamma(\cdot \mid x))\; d\mathbb{P}_k(x) < \varepsilon \tag{13}$$

and

$$W_2^2((T_\theta)_{\#}(\mathbb{P}_k \otimes \mathbb{S}), \gamma_{\#}\mathbb{P}_k) < \varepsilon. \tag{14}$$

Algorithm 1: Simultaneous Neural Optimal Transport (SimNOT)

Input: Distributions $\mathbb{P}_1, \dots, \mathbb{P}_K, \mathbb{P}^*$ and $\mathbb{S}$ accessible by samples; transport cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$; parameter $\tau > 0$; conjugate functions $\bar{\psi}, \bar{\phi}$; shared map T_θ and potentials $v_{\omega_k}, k = 1, \dots, K$; number N_T of inner iterations; batch sizes.
Output: Trained shared (stochastic) map T_θ.
repeat
  Sample batches $X_k \sim \mathbb{P}_k$, $k = 1, \dots, K$, and $Y \sim \mathbb{P}^*$;
  For each $x \in X_k$, independently sample $z_{k,x} \sim \mathbb{S}$;
  $\widehat{\mathcal{L}_v} \leftarrow -\frac{1}{K}\sum_{k=1}^{K}\Big[\frac{1}{|X_k|}\sum_{x\in X_k}\bar{\psi}(v_{\omega_k}(T_\theta(x, z_{k,x})) - \tau c(x, T_\theta(x, z_{k,x}))) + \frac{1}{|Y|}\sum_{y\in Y}\bar{\phi}(-v_{\omega_k}(y))\Big]$;
  Update $\omega_1, \dots, \omega_K$ using $\nabla_{\omega_k}\widehat{\mathcal{L}_v}$ to *maximize* $\widehat{\mathcal{L}_v}$, keeping θ fixed;
  for $n_T = 1, 2, \dots, N_T$ **do**
    Sample batches $X_k \sim \mathbb{P}_k$, $k = 1, \dots, K$;
    For each $x \in X_k$, independently sample $z_{k,x} \sim \mathbb{S}$;
    $\widehat{\mathcal{L}_T} \leftarrow -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|X_k|}\sum_{x\in X_k}\bar{\psi}(v_{\omega_k}(T_\theta(x, z_{k,x})) - \tau c(x, T_\theta(x, z_{k,x})))$;
    Update θ using $\nabla_\theta\widehat{\mathcal{L}_T}$ to *minimize* $\widehat{\mathcal{L}_T}$, keeping $\omega_1, \dots, \omega_K$ fixed;
until *converged*;

The theorem shows that any shared stochastic transport mechanism can be approximated by a single neural map, simultaneously and with a common accuracy bound for all source distributions.

Corollary 1 (Consistency of the neural parametrization)**.** *Assume the conditions of Theorem 1 and $\mathcal{Y} = [a, b]^{D_y}$. Let $\mathcal{T}_n$ be increasing classes of ReLU maps $\mathcal{X} \times [0, 1] \to \mathcal{Y}$ whose union contains every finite ReLU network with coordinatewise clipped output. Let $\mathcal{V}_m$ be increasing compact subsets of $C(\mathcal{Y})^K$ whose union is uniformly dense in $C(\mathcal{Y})^K$. Define*

$$D_{m,n} = \sup_{\mathbf{v}\in\mathcal{V}_m}\inf_{T\in\mathcal{T}_n}\mathcal{J}(\mathbf{v}, T). \tag{15}$$

Then

$$\lim_{m\to\infty}\lim_{n\to\infty} D_{m,n} = J^*. \tag{16}$$

The corollary shows that the neural parametrization is also consistent at the level of the optimization problem: as the map and potential classes become sufficiently expressive, the restricted population value converges to the exact value J^*. This statement concerns optimal values; it does not require convergence of particular network parameters. The proof is given in Appendix A.2.

We finally ask a different question. The previous theorem shows that stochastic kernels can be represented by neural maps; it does not say whether randomization is actually needed by an optimal primal solution. For quadratic cost and KL marginal penalties, an attained optimum is deterministic under a standard regularity assumption on the source densities.

Theorem 3 (Monge structure in the quadratic KL case)**.** *Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$,*

$$c(x, y) = \|x - y\|^2, \qquad \psi(t) = \phi(t) = t\log t - t + 1.$$

Assume that there exists an open set $\Omega \subseteq \mathbb{R}^d$ such that $\mathbb{P}_k(\Omega) = 1$ for all k, and the restriction of $\mathbb{P}_k$ to Ω admits a strictly positive density $p_k \in C^2(\Omega)$ with respect to d-dimensional Lebesgue measure. If the infimum in equation 8 is attained by $\big(\gamma^*, \{(\gamma_k^*)_x\}_{k=1}^K\big)$, then there exists a Borel map $T^* : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\gamma^*(\cdot \mid x) = \delta_{T^*(x)} \qquad \sigma\text{-a.e.}, \qquad \sigma = \sum_{k=1}^{K}(\gamma_k^*)_x. \tag{17}$$

Consequently,

$$\gamma_k^* = (\mathrm{id}_{\mathbb{R}^d}, T^*)_\#(\gamma_k^*)_x, \qquad k = 1, \dots, K.$$

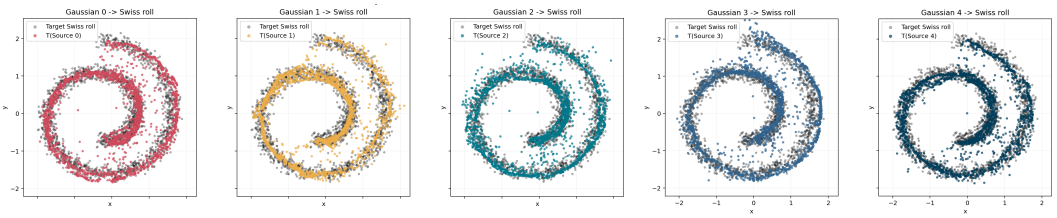

Figure 3: Simultaneous transport from five Gaussian sources to a common Swiss roll target. Each panel shows the outputs of the same learned map for one source (colored points), overlaid with target samples (gray points). The visualization uses independent samples after 100K training iterations.

5 Experiments

We evaluate SimNOT on a synthetic example and an unpaired image restoration task. In §5.1, we investigate the learned shared transformation between several Gaussian source distributions and a common Swiss roll target. In §5.2, we apply our method to CelebA images with different degradations, using a single model to restore them to a common clean image distribution.

5.1 Toy Experiment

We consider a two-dimensional example to illustrate the ability of our method to learn a shared transformation from several source distributions to a common target. The sources $\mathbb{P}_1, \dots, \mathbb{P}_5$ are Gaussian distributions with identity covariance matrices and means $(0, 0)$, $(3, 0)$, $(0, 3)$, $(-3, 0)$, and $(0, -3)$, respectively. The target $\mathbb{P}^*$ is a noisy Swiss-roll distribution. We learn a single deterministic map $T_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ with five source-specific potentials. The map receives only the input coordinates, without the source index. Further experimental details are provided in Appendix D.1.

Results. Figure 3 shows the transported samples from each source alongside samples from the common target. The shared map recovers the overall spiral structure for all five sources. The outputs differ in their spread around the spiral, and some samples lie between its branches. Thus, the experiment illustrates simultaneous approximation of the target geometry, while leaving visible discrepancies between the transported and target distributions.

5.2 Unpaired Image Restoration with Multiple Degradations

Experimental setup. We consider unpaired image restoration on CelebA at resolution 64×64. The five source distributions correspond to bicubic and bilinear downsampling, JPEG compression, Gaussian blur, and additive Gaussian noise. The common target consists of clean images. We use disjoint sets of images to construct the source distributions, the clean target, and the test set. Each source image is assigned to one degradation, and its clean counterpart is not used as a training target. Our goal is to learn a shared restoration map that processes inputs without receiving their degradation labels. Data preparation and training details are provided in Appendix D.2.

Baselines. We compare SimNOT with two UOT baselines. *Pooled UOT* learns a shared map and a single potential by treating the source distributions as an equally weighted mixture. *Conditional UOT* conditions both the map and the potential on the degradation label. For blind restoration, this label is predicted by a separately trained degradation classifier which uses the same five degradation types and fixed parameters as restoration training. All models use the same data split and are evaluated using their final exponential moving average (EMA) checkpoints after 100K training steps.

Evaluation. We assess distribution matching using FID and reconstruction accuracy using PSNR against the corresponding clean images. For the original degradation settings, we compute FID separately for each source and report the mean and maximum across the five degradations.

Results on training degradations. Table 1 reports results on held-out images with the degradation parameters used during training. SimNOT outperforms Pooled UOT in both mean and maximum source-specific FID and in PSNR. Conditional UOT achieves the best results in this setting. The

Method	Mean FID ↓	Max FID ↓	PSNR ↑
Degraded input	87.60	178.94	25.19
Pooled UOT	9.96	14.39	27.51
Conditional UOT + classifier	**6.39**	**8.33**	**28.21**
SimNOT (ours)	<u>8.23</u>	<u>11.88</u>	<u>27.85</u>

Table 1: CelebA restoration on held-out images with the training degradation parameters. Each of the five degradations is applied to the same full set of test images. Mean and Max FID aggregate the five source-specific FIDs. PSNR (dB) is averaged over images and degradation conditions. The best and second-best results among restoration methods are bold and underlined, respectively.

	Bilinear $\times 3$			Bilinear $\times 5$		
Method	FID ↓	LPIPS ↓	PSNR ↑	FID ↓	LPIPS ↓	PSNR ↑
Degraded input	92.82	0.1821	25.02	317.44	0.3717	20.93
Pooled UOT	27.71	0.0717	24.23	99.22	0.2342	20.51
Conditional UOT + classifier	35.16	0.1185	24.35	80.87	0.2670	20.80
SimNOT (ours)	**18.44**	**0.0635**	**25.03**	**32.15**	**0.1134**	**20.93**
Conditional UOT + known family	21.42	0.0637	23.99	77.23	0.1983	20.69

Table 2: Generalization to unseen bilinear downsampling factors. Models trained with factor $\times 4$ are evaluated at factors $\times 3$ and $\times 5$ without retraining. Each setting uses all $20{,}259$ test images. The last row supplies the bilinear family label directly to Conditional UOT. Bold indicates the best result among restoration methods.

degradation classifier attains 100% accuracy, so using its predictions yields the same results as supplying the true degradation labels.

Generalization to unseen degradation strengths. We further evaluate the trained models on bilinear downsampling with factors $\times 3$ and $\times 5$, while training uses factor $\times 4$. All models, including the classifier, are applied without retraining. We additionally report LPIPS to assess perceptual similarity to the clean reference images.

As shown in Table 2, SimNOT achieves the lowest FID and LPIPS and the highest PSNR among the restoration methods at both factors. The classifier fails to identify the bilinear family in either setting. To distinguish classification errors from limitations of the conditional model, we also evaluate Conditional UOT with the known bilinear label. It improves FID and LPIPS, but SimNOT retains an advantage, particularly at factor $\times 5$. Experiments with altered bicubic, JPEG, blur, and noise parameters are reported in Appendix C. While performance of SimNOT varies across degradations, it preserves perceptual similarity more consistently under shifts in degradation parameters.

6 Conclusion

We introduce SimNOT, a neural solver that brings simultaneous OT to the continuous, sample-based setting. Our method learns a single transport map for multiple source distributions, with a separate distribution-matching objective for each source. We develop a divergence-based unbalanced formulation, derive its exact semi-dual representation, and establish neural approximation guarantees. Experiments on synthetic data and image restoration demonstrate that SimNOT learns shared transformations from unpaired samples and handles multiple degradations without requiring their labels at inference time. Our work extends the scope of neural OT from pairwise transport to simultaneous learning across multiple sources, providing a principled framework for this broader class of transformation problems. We discuss limitations and future research directions in Appendix E.

Reproducibility statement

Experimental details, including data preparation, architectures, hyperparameters, and evaluation protocols, are provided in Appendix D. Our code and instructions for reproducing the experiments will be made publicly available.

AI USE STATEMENT

AI tools were used to assist with polishing and drafting the text, checking proofs, and supporting the implementation and analysis of experiments. All AI-assisted work was reviewed and verified by the authors, who take full responsibility for the final content of this work.

REFERENCES

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

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.

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

Lénaïc Chizat. *Unbalanced optimal transport: Models, numerical methods, applications*. PhD thesis, Université Paris sciences et lettres, 2017.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Generative modeling through the semi-dual formulation of unbalanced optimal transport. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Chanjun Chun, Young Han Lee, Geon Woo Lee, Moongu Jeon, and Hong Kook Kim. Non-parallel voice conversion using cycle-consistent adversarial networks with self-supervised representations. In *2023 IEEE 20th consumer communications & networking conference (CCNC)*, pp. 931–932. IEEE, 2023.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, volume 26, pp. 2292–2300, 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Luca Eyring, Dominik Klein, Théo Uscidda, Giovanni Palla, Niki Kilbertus, Zeynep Akata, and Fabian Theis. Unbalancedness in neural monge maps improves unpaired domain translation. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiaojiao Fan, Shu Liu, Shaojun Ma, Yongxin Chen, and Hao-Min Zhou. Scalable computation of monge maps with general costs. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2021.

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. ISSN 2835-8856. URL https://openreview.net/forum?id=2mZSlQscj3. Featured Certification.

Milena Gazdieva, Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Extremal domain translation with neural optimal transport. In *Advances in Neural Information Processing Systems*, 2023.

Milena Gazdieva, Arip Asadulaev, Evgeny Burnaev, and Alexander Korotin. Light unbalanced optimal transport. *Advances in Neural Information Processing Systems*, 37:93907–93938, 2024.

Milena Gazdieva, Jaemoo Choi, Alexander Kolesov, Jaewoong Choi, Petr Mokrov, and Aleksandr Korotin. Robust barycenter estimation using semi-unbalanced neural optimal transport. In *International Conference on Learning Representations*, volume 2025, pp. 74233–74262, 2025.

Dominik Klein, Théo Uscidda, Fabian Theis, and Marco Cuturi. Genot: A neural optimal transport framework for single cell genomics. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Evgeny Burnaev, and Alexander Korotin. Estimating barycenters of distributions with neural optimal transport. In *Forty-first International Conference on Machine Learning*, 2024a.

Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Anastasis Kratsios, Evgeny Burnaev, and Alexander Korotin. Energy-guided continuous entropic barycenter estimation for general costs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Kernel neural optimal transport. In *International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=Zuc_MHtUma4.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=d8CBRlWNkqH.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Li_All-in-One_Image_Restoration_for_Unknown_Corruption_CVPR_2022_paper.html.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3): 969–1117, 2018.

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.

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

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Khan. Promptir: Prompting for all-in-one image restoration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*, 2022.

Xiaole Tang, Xiang Gu, Xiaoyi He, Xin Hu, and Jian Sun. Degradation-aware residual-conditioned optimal transport for unified image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6764–6779, 2025. doi: 10.1109/TPAMI.2025.3562211. URL https://doi.org/10.1109/TPAMI.2025.3562211.

Xiaole Tang, Xiaoyi He, Jiayi Xu, Xiang Gu, and Jian Sun. Learning continuous Wasserstein barycenter space for generalized all-in-one image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3669121. URL https://doi.org/10.1109/TPAMI.2026.3669121.

Cédric Villani. *Optimal transport: old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Ruodu Wang and Zhenyuan Zhang. Simultaneous optimal transport. *Transactions of the American Mathematical Society*, 378(8):5845–5898, 2025. doi: 10.1090/tran/9393. URL https://doi.org/10.1090/tran/9393.

Karren D Yang and Caroline Uhler. Scalable unbalanced optimal transport using generative adversarial networks. In *International Conference on Learning Representations*, 2018.

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.

A PROOFS OF THE THEORETICAL RESULTS

A.1 PROOF OF THE SEMI-DUAL FORMULATION

Put

$$\overline{\mathbb{P}} = \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}_k, \qquad r_k = \frac{d\mathbb{P}_k}{d\overline{\mathbb{P}}}.$$

We choose Borel versions such that $0 \le r_k \le K$ and $\sum_k r_k = K\ \overline{\mathbb{P}}$-almost everywhere.

For $h = \psi$ or $h = \phi$, we use the standard variational identity

$$D_h(\mu\|\nu) = \sup_{f\in C_b(E)}\left\{\int_E f\,d\mu - \int_E \bar{h}(f)\,d\nu\right\}$$

for finite nonnegative Borel measures on a Polish space E. Under the assumptions of Theorem 1, these divergences are jointly lower semicontinuous for narrow convergence and satisfy the data-processing inequality.

For a probability kernel $\gamma(\cdot \mid x)$, define the probability measure

$$m_k^\gamma(dx, dy) = \mathbb{P}_k(dx)\,\gamma(dy \mid x).$$

Consider the convex relaxation

$$R = \inf_{\gamma,\,\pi_1,\dots,\pi_K\ge 0}\frac{1}{K}\sum_{k=1}^{K}\left[\int c\,d\pi_k + D_\psi(\pi_k\|m_k^\gamma) + D_\phi((\pi_k)_y\|\mathbb{P}^*)\right]. \tag{18}$$

Every admissible point of equation 8 is admissible here: if

$$\pi_k(dx, dy) = \alpha_k(dx)\,\gamma(dy \mid x),$$

then

$$D_\psi(\pi_k\|m_k^\gamma) = D_\psi(\alpha_k\|\mathbb{P}_k).$$

Hence $R \le J^*$.

The relaxation allows the density of π_k with respect to m_k^γ to depend on both x and y. The following lemma shows that this additional freedom does not change the infimum.

Lemma 1 (No relaxation gap)**.** *Under the assumptions of Theorem 1,*

$$R = J^*.$$

Proof. Fix an admissible point of equation 18 with finite value and write

$$s_k = \frac{d\pi_k}{dm_k^\gamma}.$$

We first reduce to bounded finite-valued densities supported on a common compact rectangle. By tightness, choose increasing compact rectangles $E_n = C_n \times D_n$ such that $m_k^\gamma(E_n^c) \to 0$ for every k, and set

$$s_{k,n} = 2^{-n}\,\lfloor 2^n \min\{s_k, n\}\rfloor\,\mathbf{1}_{E_n}.$$

Then $0 \le s_{k,n} \uparrow s_k$ m_k^γ-almost everywhere. Monotone convergence gives convergence of the transport costs. Since for a nonnegative convex function h and $0 \le a \le b$,

$$h(a) \le h(0) + h(b),$$

dominated convergence gives convergence of the source divergence terms. The target measures $(s_{k,n}m_k^\gamma)_y$ increase to $(\pi_k)_y$, and the same bound applied to their densities with respect to $\mathbb{P}^*$ gives convergence of the target divergences. It is therefore enough to treat finite-valued s_k supported on one compact rectangle $C \times D$.

Fix $\delta > 0$. Uniform continuity of c on $C \times D$ gives a finite Borel partition

$$D = B_1 \sqcup \cdots \sqcup B_L$$

such that

$$\sup_{x\in C}\sup_{y,y'\in B_\ell}|c(x,y)-c(x,y')|\le\delta$$

for every ℓ. Cells with $\mathbb{P}^*(B_\ell)=0$ carry no transported mass and may be ignored. For the remaining cells, put

$$\nu_\ell=\frac{\mathbb{P}^*|_{B_\ell}}{\mathbb{P}^*(B_\ell)},\qquad \bar{c}_\ell(x)=\int_{B_\ell}c(x,y)\,d\nu_\ell(y).$$

For fixed x, sample $y\sim\gamma(\cdot\mid x)$ and record the finite action

$$(\ell,s_1(x,y),\dots,s_K(x,y))$$

when $y\in B_\ell$; points carrying no transported mass are assigned one additional zero action. This gives a randomized rule with a finite action space. Since $\overline{\mathbb{P}}$ is atomless, it can be approximated by measurable deterministic rules while preserving, up to an arbitrarily small error, the finitely many integrals corresponding to

$$r_k(x)\psi(z_k),\qquad r_k(x)z_k\mathbf{1}_{\{\ell=j\}},\qquad r_k(x)z_k\bar{c}_\ell(x).$$

One may obtain this directly by approximating these payoffs by simple functions on a common finite partition of C and splitting each atomless partition cell according to the required action probabilities.

Let

$$x\mapsto(\ell_n(x),z_{1,n}(x),\dots,z_{K,n}(x))$$

be such deterministic rules. Define

$$\alpha_{k,n}(dx)=z_{k,n}(x)\,\mathbb{P}_k(dx)$$

on C and set $\alpha_{k,n}=0$ outside C. At a nonzero action, define the common kernel by

$$\gamma_n(dy\mid x)=\nu_{\ell_n(x)}(dy),$$

and choose it arbitrarily at the zero action and outside C.

By construction, the source divergence terms converge to $D_\psi(\pi_k\|m_k^\gamma)$. The target measures have the form

$$(\alpha_{k,n}(dx)\gamma_n(dy\mid x))_y=\sum_{\ell=1}^{L}b_{k\ell,n}\nu_\ell,$$

where

$$b_{k\ell,n}\longrightarrow b_{k\ell}:=(\pi_k)_y(B_\ell).$$

If $\rho_k=d(\pi_k)_y/d\mathbb{P}^*$, then Jensen's inequality gives

$$\mathbb{P}^*(B_\ell)\phi\left(\frac{b_{k\ell}}{\mathbb{P}^*(B_\ell)}\right)\le\int_{B_\ell}\phi(\rho_k)\,d\mathbb{P}^*.$$

Hence the limiting target divergence is no larger than $D_\phi((\pi_k)_y\|\mathbb{P}^*)$.

Finally, the limiting transport cost differs from $\int c\,d\pi_k$ by at most

$$\delta\,\pi_k(\mathcal{X}\times\mathcal{Y}).$$

Letting the purification error tend to zero, then $\delta\downarrow0$, and finally removing the initial finite-valued compact approximation gives $J^*\le R$. Together with $R\le J^*$, this proves the lemma. □

Proof of Theorem 1. By Lemma 1, it remains to derive the dual of the convex problem equation 18.

We first explain the argument for a compact target. Write a shared kernel as

$$q(dx,dy)=\overline{\mathbb{P}}(dx)\,\gamma(dy\mid x)$$

and set

$$Q=\left\{q\in\mathcal{P}(\mathcal{X}\times\mathcal{Y}):q_x=\overline{\mathbb{P}}\right\}.$$

Then

$$m_k^\gamma=r_kq.$$

The set Q is convex and narrowly compact. Moreover, $q \mapsto r_k q$ is narrowly continuous on Q: approximate the bounded Borel function r_k in $L^1(\overline{\mathbb{P}})$ by bounded continuous functions and use that every $q \in Q$ has the same x-marginal.

For finite nonnegative measures π_k on $\mathcal{X} \times \mathcal{Y}$ and β_k on $\mathcal{Y}$, define

$$E(q, \boldsymbol{\pi}, \boldsymbol{\beta}) = \sum_{k=1}^{K} \left[\int c \, d\pi_k + D_\psi(\pi_k \| r_k q) + D_\phi(\beta_k \| \mathbb{P}^*) \right].$$

Its sublevel sets are narrowly compact. Indeed, superlinearity of ψ gives uniform bounds on the masses and tails of π_k relative to the uniformly tight family $\{r_k q : q \in Q\}$; the same argument applies to β_k relative to $\mathbb{P}^*$. Lower semicontinuity follows from lower semicontinuity of the cost and divergences and from continuity of $q \mapsto r_k q$.

For signed finite measures $\eta_1, \ldots, \eta_K$ on $\mathcal{Y}$, define

$$W(\boldsymbol{\eta}) = \inf_{\substack{q \in Q,\ \pi_k, \beta_k \geq 0 \\ (\pi_k)_y - \beta_k = \eta_k}} E(q, \boldsymbol{\pi}, \boldsymbol{\beta}).$$

The function W is proper, convex and lower semicontinuous in the weak topology induced by $C(\mathcal{Y})^K$. Its finite infima are attained by compactness of the energy sublevels, and

$$W(0) = KR.$$

Fenchel–Moreau therefore gives

$$KR = \sup_{\mathbf{v} \in C(\mathcal{Y})^K} \{-W^*(\mathbf{v})\}.$$

We compute the conjugate pointwise. For fixed q and $v \in C(\mathcal{Y})$,

$$\sup_{\pi \geq 0} \left\{ \int v(y) \, d\pi - \int c \, d\pi - D_\psi(\pi \| r_k q) \right\} = \int \bar{\psi}(v(y) - c(x, y)) \, d(r_k q)(x, y).$$

To justify the interchange of supremum and integration, note that $v - c$ is bounded above. Superlinearity of ψ therefore bounds all pointwise maximizers in a common compact interval. Using a countable dense subset of the effective domain (including its finite endpoints), one obtains measurable finite-valued near-maximizers. Choosing them at least as good as $t = 0$ also gives a uniform bound on $c(x, y)t$, so the corresponding measures have finite transport cost. The resulting integrals converge to the pointwise conjugate. Similarly,

$$\sup_{\beta \geq 0} \left\{ -\int v \, d\beta - D_\phi(\beta \| \mathbb{P}^*) \right\} = \int \bar{\phi}(-v) \, d\mathbb{P}^*.$$

Substitution into the Fenchel–Moreau formula yields

$$R = \sup_{\mathbf{v} \in C(\mathcal{Y})^K} \inf_{\gamma} \mathcal{J}(\mathbf{v}, \gamma)$$

when $\mathcal{Y}$ is compact.

Now suppose that $\mathcal{Y}$ is noncompact. Since it is a closed subset of Euclidean space, its one-point compactification

$$\widehat{\mathcal{Y}} = \mathcal{Y} \cup \{\infty\}$$

is a compact metric space. Extend $\mathbb{P}^*$ by zero at ∞ and set

$$\widehat{c}(x, y) = c(x, y) \quad \text{for } y \in \mathcal{Y}, \qquad \widehat{c}(x, \infty) = 0.$$

The extended cost is lower semicontinuous because $c \geq 0$. Compactification does not change the relaxed value. Indeed, every finite-value candidate has $\pi_k(\mathcal{X} \times \{\infty\}) = 0$ because $(\pi_k)_y \ll \mathbb{P}^*$. If the reference kernel places mass at ∞, send this mass to an arbitrary fixed point $y_0 \in \mathcal{Y}$. The transport plans and their target marginals are unchanged, while the data-processing inequality shows that the source divergence cannot increase.

Applying the compact argument on $\widehat{\mathcal{Y}}$ and then restricting the potentials and kernels back to $\mathcal{Y}$ gives

$$R \le \sup_{\mathbf{v}\in C_b(\mathcal{Y})^K} \inf_{\gamma} \mathcal{J}(\mathbf{v},\gamma).$$

Indeed, restrictions of functions in $C(\widehat{\mathcal{Y}})$ belong to $C_b(\mathcal{Y})$, while kernels supported on $\mathcal{Y}$ form a subclass of the compactified kernels. For the reverse bound, let $v_k \in C_b(\mathcal{Y})$ and use the pointwise Fenchel inequalities

$$\int c\, d\pi_k + D_\psi(\pi_k \| m_k^\gamma) \ge \int v_k\, d(\pi_k)_y - \int \bar{\psi}(v_k - c)\, dm_k^\gamma$$

and

$$D_\phi((\pi_k)_y \| \mathbb{P}^*) \ge -\int v_k\, d(\pi_k)_y - \int \bar{\phi}(-v_k)\, d\mathbb{P}^*.$$

After summing and taking infima, this gives

$$\sup_{\mathbf{v}\in C_b(\mathcal{Y})^K} \inf_{\gamma} \mathcal{J}(\mathbf{v},\gamma) \le R.$$

Together with the compactified lower bound and Lemma 1,

$$J^* = \sup_{\mathbf{v}\in C_b(\mathcal{Y})^K} \inf_{\gamma} \mathcal{J}(\mathbf{v},\gamma).$$

It remains to remove the inner randomization. Fix $\mathbf{v} \in C_b(\mathcal{Y})^K$ and define

$$H_{\mathbf{v}}(x,y) = \frac{1}{K}\sum_{k=1}^{K} r_k(x)\bar{\psi}(v_k(y) - c(x,y)).$$

For each fixed x, the function $y \mapsto H_{\mathbf{v}}(x,y)$ is continuous, and $H_{\mathbf{v}}$ is bounded because

$$-\psi(0) \le \bar{\psi}(v_k(y) - c(x,y)) \le \bar{\psi}(\|v_k\|_\infty).$$

Choose a countable dense set $\{y_j : j \ge 1\} \subseteq \mathcal{Y}$. Then

$$h(x) = \sup_{y\in\mathcal{Y}} H_{\mathbf{v}}(x,y) = \sup_{j\ge 1} H_{\mathbf{v}}(x,y_j)$$

is Borel. For every $\varepsilon > 0$, choosing the first y_j with

$$H_{\mathbf{v}}(x,y_j) > h(x) - \varepsilon$$

defines a Borel map $T_\varepsilon : \mathcal{X} \to \mathcal{Y}$. Hence

$$\sup_{\gamma} \int H_{\mathbf{v}}(x,y)\, \gamma(dy \mid x)\, d\overline{\mathbb{P}}(x) = \sup_{T} \int H_{\mathbf{v}}(x,T(x))\, d\overline{\mathbb{P}}(x),$$

where the second supremum is over Borel maps. Since the target term in $\mathcal{J}$ does not depend on the kernel,

$$\inf_{\gamma} \mathcal{J}(\mathbf{v},\gamma) = \inf_{T} \mathcal{J}(\mathbf{v},T).$$

This proves Theorem 1. □

A.2 Proof of the approximation results

Proof of Theorem 2. By the randomization lemma, there exists a measurable map

$$T : \mathcal{X} \times [0,1] \to \mathcal{Y}$$

such that

$$(T(x,\cdot))_{\#}\mathbb{S} = \gamma(\cdot \mid x).$$

Put

$$\overline{\mathbb{P}} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{P}_k, \qquad \mu = \overline{\mathbb{P}} \otimes \mathbb{S}.$$

Since $\mathcal{Y} = [a, b]^{D_y}$ is bounded, $T \in L^2(\mu; \mathbb{R}^{D_y})$.

We claim that T can be approximated arbitrarily well in $L^2(\mu)$ by finite ReLU networks with output in $\mathcal{Y}$. By Lusin's theorem, for every $\eta > 0$ there is a compact set $K_\eta \subseteq \mathcal{X} \times [0, 1]$ such that $T|_{K_\eta}$ is continuous and $\mu(K_\eta^c) < \eta$. The usual ReLU universal approximation theorem gives a finite network that approximates T uniformly on K_η. Clipping each output coordinate to $[a, b]$ is itself a ReLU operation and cannot increase the error to a point of $\mathcal{Y}$. Since both T and the clipped network are uniformly bounded, the contribution of K_η^c to the L^2 error tends to zero with η. Thus, for any $\delta > 0$, we can choose T_θ such that

$$\int \|T_\theta(x, z) - T(x, z)\|^2 \, d\overline{\mathbb{P}}(x) \, d\mathbb{S}(z) < \delta.$$

Because $\mathbb{P}_k \leq K\overline{\mathbb{P}}$,

$$\int \|T_\theta(x, z) - T(x, z)\|^2 \, d\mathbb{P}_k(x) \, d\mathbb{S}(z) < K\delta$$

for every k. Choose $\delta < \varepsilon/K$.

For fixed x, the pair

$$(T_\theta(x, Z), T(x, Z)), \qquad Z \sim \mathbb{S},$$

is a coupling of $\gamma_\theta(\cdot \mid x)$ and $\gamma(\cdot \mid x)$. Therefore

$$W_2^2(\gamma_\theta(\cdot \mid x), \gamma(\cdot \mid x)) \leq \mathbb{E}\|T_\theta(x, Z) - T(x, Z)\|^2.$$

Integrating with respect to $\mathbb{P}_k$ proves equation 13. Sampling additionally $X \sim \mathbb{P}_k$ gives a coupling of the two output laws with the same quadratic cost, which proves equation 14. □

Proof of Corollary 1. For fixed $\mathbf{v} \in C(\mathcal{Y})^K$, set

$$d_n(\mathbf{v}) = \inf_{T \in \mathcal{T}_n} \mathcal{J}(\mathbf{v}, T), \qquad d(\mathbf{v}) = \inf_T \mathcal{J}(\mathbf{v}, T),$$

where the second infimum is over all measurable maps $T : \mathcal{X} \times [0, 1] \to \mathcal{Y}$.

The proof of Theorem 2 shows that every such measurable map can be approximated in $L^2(\overline{\mathbb{P}} \otimes \mathbb{S})$ by maps from $\bigcup_n \mathcal{T}_n$. Hence the approximating maps converge in probability under every $\mathbb{P}_k \otimes \mathbb{S}$. Since c and the potentials are continuous and

$$-\psi(0) \leq \bar{\psi}(v_k(y) - c(x, y)) \leq \bar{\psi}(\|v_k\|_\infty),$$

the corresponding integrands are uniformly bounded. It follows that the expectations converge, and therefore

$$d_n(\mathbf{v}) \downarrow d(\mathbf{v}).$$

Fix m. Compactness of $\mathcal{V}_m$ in the uniform norm gives a common bound on its potentials. Superlinearity of ψ implies that $\bar{\psi}$ is Lipschitz on every interval $(-\infty, B]$, while $\bar{\phi}$ is Lipschitz on bounded intervals. Consequently, d_n and d are uniformly Lipschitz on $\mathcal{V}_m$. Since $d_n \downarrow d$ and $\mathcal{V}_m$ is compact, Dini's theorem gives uniform convergence and therefore

$$\lim_{n \to \infty} D_{m,n} = \sup_{\mathbf{v} \in \mathcal{V}_m} d(\mathbf{v}).$$

Finally, $\bigcup_m \mathcal{V}_m$ is uniformly dense in $C(\mathcal{Y})^K$ and d is continuous. Hence

$$\lim_{m \to \infty} \sup_{\mathbf{v} \in \mathcal{V}_m} d(\mathbf{v}) = \sup_{\mathbf{v} \in C(\mathcal{Y})^K} d(\mathbf{v}).$$

Because $\mathcal{Y}$ is compact, $C(\mathcal{Y}) = C_b(\mathcal{Y})$, and Theorem 1 identifies the last quantity with J^*. This proves equation 16. □

A.3 PROOF OF THE DETERMINISTIC STRUCTURE THEOREM

Proof of Theorem 3. Write

$$\alpha_k = \alpha_k^*, \qquad M = \sum_{k=1}^K \mathbb{P}_k, \qquad \sigma = \sum_{k=1}^K \alpha_k.$$

Finite KL divergence gives

$$\sigma \ll M \ll \mathcal{L}^d.$$

If $\sigma = 0$, the conclusion is vacuous. Assume $\sigma \neq 0$ and define on Ω

$$a_k(x) = \frac{p_k(x)}{\sum_j p_j(x)}, \qquad \lambda_k(x) = \frac{d\alpha_k}{d\sigma}(x).$$

Then $a_k > 0$, $\sum_k a_k = 1$, and $\sum_k \lambda_k = 1$ σ-almost everywhere. The functions a_k are C^2 on Ω. Let $m = d\sigma/dM$. Since

$$\frac{d\alpha_k}{d\mathbb{P}_k} = \frac{m\lambda_k}{a_k},$$

a direct calculation gives

$$\sum_{k=1}^K \mathrm{KL}(\alpha_k \| \mathbb{P}_k) = \mathrm{KL}(\sigma \| M) + \int_\Omega \sum_{k=1}^K \lambda_k(x) \log \frac{\lambda_k(x)}{a_k(x)} \, d\sigma(x). \tag{19}$$

Let

$$\Delta_K = \left\{ z \in [0,1]^K : \sum_{k=1}^K z_k = 1 \right\}$$

be the probability simplex, and write $\lambda(x) = (\lambda_1(x), \dots, \lambda_K(x))$. Define the finite measure

$$\zeta(dx, dy, dz) = \sigma(dx)\, \gamma^*(dy \mid x)\, \delta_{\lambda(x)}(dz)$$

on $\Omega \times \mathbb{R}^d \times \Delta_K$, and let ξ be its (y, z)-marginal. Consider the cost

$$\widetilde{c}(x; y, z) = \|x - y\|^2 + \sum_{k=1}^K z_k \log \frac{z_k}{a_k(x)}, \qquad 0 \log 0 := 0.$$

We first show that ζ is an optimal transport plan between σ and ξ for this cost.

Let $\widetilde{\zeta} \in \Pi(\sigma, \xi)$ and let η be its (x, y)-marginal. Write

$$\eta(dx, dy) = \sigma(dx)\, \widetilde{\gamma}(dy \mid x)$$

and define

$$\widetilde{\pi}_k(dx, dy) = \int_{\Delta_K} z_k \, \widetilde{\zeta}(dx, dy, dz).$$

If

$$\overline{z}_k(x, y) = \mathbb{E}_{\widetilde{\zeta}}[z_k \mid x, y],$$

then

$$\widetilde{\pi}_k(dx, dy) = \overline{z}_k(x, y)\, \eta(dx, dy), \qquad \sum_k \overline{z}_k = 1.$$

Because the (y, z)-marginal of $\widetilde{\zeta}$ is fixed,

$$(\widetilde{\pi}_k)_y(B) = \int_{B \times \Delta_K} z_k \, d\xi(y, z),$$

so these target marginals are the same as those of the original optimal plans.

Using the same density calculation as in equation 19, now on $\mathcal{X} \times \mathbb{R}^d$, we obtain

$$\begin{aligned}
&\sum_{k=1}^{K} \mathrm{KL}(\widetilde{\pi}_k \| \mathbb{P}_k(dx)\, \widetilde{\gamma}(dy \mid x)) \\
&\qquad = \mathrm{KL}(\sigma \| M) + \int \sum_{k=1}^{K} \overline{z}_k(x,y) \log \frac{\overline{z}_k(x,y)}{a_k(x)}\, d\eta(x,y) \\
&\qquad \le \mathrm{KL}(\sigma \| M) + \int \sum_{k=1}^{K} z_k \log \frac{z_k}{a_k(x)}\, d\widetilde{\zeta}(x,y,z),
\end{aligned}$$

where the inequality is conditional Jensen. Moreover,

$$\sum_{k=1}^{K} \int \|x-y\|^2\, d\widetilde{\pi}_k(x,y) = \int \|x-y\|^2\, d\widetilde{\zeta}(x,y,z).$$

Thus, if $\widetilde{\zeta}$ had strictly smaller $\widetilde{c}$-cost than ζ, the measures $\widetilde{\pi}_k$ together with the common kernel $\widetilde{\gamma}$ would give an admissible point of the relaxed problem equation 18 with value strictly below J^*. This contradicts Lemma 1. Hence ζ is optimal.

It remains to show that its y-coordinate is uniquely determined by x. Cover Ω by countably many open balls B_j with $\overline{B}_j \subset \Omega$, and choose larger open balls U_j such that

$$\overline{B}_j \subset U_j, \qquad \overline{U}_j \subset \Omega.$$

For $\ell \ge 1$, let

$$D_\ell = \{y \in \mathbb{R}^d : \|y\| \le \ell\}.$$

The restriction of ζ to

$$\overline{B}_j \times D_\ell \times \Delta_K$$

is optimal between its own marginals: otherwise it could be replaced by a cheaper coupling without changing the full marginals of ζ.

On $\overline{U}_j \times D_\ell \times \Delta_K$, the cost $\widetilde{c}$ is continuous and continuously differentiable in x, with bounded x-gradient. Compact Kantorovich duality gives dual potentials. Taking the cost transform on the source side gives a source potential $u_{j\ell}$ that is Lipschitz on U_j, hence differentiable Lebesgue-almost everywhere.

At a differentiability point where complementary slackness holds,

$$\nabla u_{j\ell}(x) = 2(x-y) - \sum_{k=1}^{K} z_k \nabla \log a_k(x).$$

For ζ we have $z = \lambda(x)$ almost everywhere, so the conditional measure can charge only

$$y = x - \frac{1}{2}\left[\nabla u_{j\ell}(x) + \sum_{k=1}^{K} \lambda_k(x) \nabla \log a_k(x)\right].$$

Since $\sigma \ll \mathcal{L}^d$, after removing one σ-null set for the countable family (j,ℓ), this conclusion holds whenever the corresponding restricted conditional measure is nonzero. The sets D_ℓ increase to $\mathbb{R}^d$, and their restrictions are consistent. Therefore $\gamma^*(\cdot \mid x)$ is a Dirac mass for σ-almost every x.

A Borel probability kernel that is almost everywhere a Dirac mass induced, after modification on a null set, by a Borel map. Hence there exists $T^* : \mathbb{R}^d \to \mathbb{R}^d$ satisfying equation 17, and the formula for the optimal plans follows immediately. □

B Extended Discussion of Related Work

OT barycenters. Barycenter methods also relate several distributions through OT. Continuous neural solvers estimate a common distribution by minimizing an average of transport costs (Fan et al., 2021;

Kolesov et al., 2024b;a); semi-unbalanced extensions allow robust barycenter estimation (Gazdieva et al., 2025). In these problems, the common distribution is an unknown to be optimized. In our setting, the target distribution is prescribed and accessible through samples. Moreover, the simultaneous constraint requires a shared transformation, whereas the standard barycenter problem allows separate transport plans for different input distributions.

All-in-one image restoration. Learning a single model for multiple degradations is the aim of all-in-one image restoration. AirNet (Li et al., 2022) learns degradation representations, while PromptIR (Potlapalli et al., 2023) uses learned prompts to adapt restoration to the input. Their restoration networks are trained with paired degraded and clean images. OT-based approaches also address this task: DA-RCOT (Tang et al., 2025) considers both unpaired and paired restoration, using transport residuals to inform the cost and condition the restoration map. Our focus is on imposing separate distribution-matching objectives for the individual sources within a simultaneous formulation. Another closely related approach, BaryIR (Tang et al., 2026), learns a shared map and multiple potentials to construct a Wasserstein barycenter representation of degraded features. It combines this representation with residual features and uses paired supervision for restoration. SimNOT instead learns transport to a prescribed target distribution from unpaired samples. Thus, the key distinction from BaryIR lies in the transport problem and supervision, rather than in sharing a map across degradations.

C Additional Experimental Results

C.1 Results for Individual Degradations

Table 3 provides the source-specific results underlying Table 1. It shows that SimNOT achieves lower FID and higher PSNR than Pooled UOT for JPEG compression, Gaussian blur, and Gaussian noise, although Pooled UOT performs slightly better for the downsampling degradations. Besides, SimNOT obtains the best FID and PSNR for Gaussian noise and the best PSNR for JPEG compression, while conditional UOT performs best in FID for the remaining four degradations.

Method	Bicubic	Bilinear	JPEG	Blur	Noise
			FID ↓		
Degraded input	124.86	178.94	31.19	46.47	56.54
Pooled UOT	<u>10.96</u>	<u>8.54</u>	9.15	6.75	14.39
Conditional UOT + classifier	**8.33**	**8.18**	**5.92**	**3.48**	<u>6.02</u>
SimNOT (ours)	11.28	11.88	<u>6.59</u>	<u>6.27</u>	**5.14**
			PSNR (dB) ↑		
Degraded input	22.89	23.02	28.50	25.22	26.32
Pooled UOT	<u>22.96</u>	**23.74**	29.22	29.84	31.78
Conditional UOT + classifier	**22.99**	<u>23.72</u>	<u>29.55</u>	**31.34**	<u>33.42</u>
SimNOT (ours)	22.70	23.26	**29.59**	<u>29.86</u>	**33.85**

Table 3: Source-specific results on the original degradation settings. The *evaluation protocol* is described in Appendix D.2. The best and second-best results among restoration methods are bold and underlined, respectively. Rankings use unrounded values.

C.2 Generalization to Unseen Degradation Strengths

We extend the bilinear experiments in Table 2 to altered bicubic downsampling factors, JPEG quality levels, blur standard deviations, and noise standard deviations. Tables 4 and 5 report results for milder and stronger degradations, respectively. Each setting uses all $20{,}259$ test images and the same trained models, without fine-tuning.

The results show that SimNOT achieves lower LPIPS than classifier-routed Conditional UOT in $9/10$ conditions and Pooled UOT in $7/10$. SimNOT also has the highest mean PSNR among the blind methods, although Pooled UOT achieves a slightly lower mean FID. The known-family reference has

Method	FID ↓	LPIPS ↓	PSNR ↑
Bicubic $\times 3$ (training: $\times 4$)			
Degraded input	61.46	0.16835	25.24
Pooled UOT	<u>21.06</u>	<u>0.07670</u>	**23.10**
Conditional UOT + classifier	113.85	0.20161	20.25
SimNOT (ours)	**18.72**	**0.06642**	<u>22.49</u>
Conditional UOT + known family	9.27	0.04598	23.74
JPEG quality 30 (training: 25)			
Degraded input	27.22	0.03445	29.11
Pooled UOT	8.30	0.01726	29.74
Conditional UOT + classifier	**5.48**	**0.01546**	<u>30.00</u>
SimNOT (ours)	<u>6.40</u>	<u>0.01598</u>	**30.02**
Conditional UOT + known family	5.48	0.01546	30.00
Gaussian blur $\sigma = 1.25$ (training: 1.5)			
Degraded input	31.02	0.19583	26.62
Pooled UOT	<u>6.80</u>	**0.01804**	**30.26**
Conditional UOT + classifier	**6.61**	0.03509	<u>29.09</u>
SimNOT (ours)	8.95	<u>0.03057</u>	28.41
Conditional UOT + known family	6.61	0.03508	29.09
Gaussian noise $\sigma = 0.04$ (training: 0.05)			
Degraded input	48.45	0.02418	28.20
Pooled UOT	6.60	0.00867	32.85
Conditional UOT + classifier	**2.85**	<u>0.00543</u>	<u>34.60</u>
SimNOT (ours)	<u>3.17</u>	**0.00510**	**35.08**
Conditional UOT + known family	2.85	0.00543	34.60

Table 4: Generalization to milder degradations. All models are evaluated without retraining on 20,259 images per setting. The best and second-best results among blind restoration methods are bold and underlined, respectively. The known-family variant receives the true degradation family and is reported separately as a reference. Rankings use unrounded values.

better mean scores in all three metrics. Thus, among the blind methods, SimNOT's most consistent advantage is in perceptual similarity measured by LPIPS.

C.3 Degradation Classification Results

The degradation classifier achieves 100% accuracy on the test set with the original degradation parameters. Table 6 reports its accuracy under parameter shifts, measured against the known degradation family.

Classification remains accurate for the tested JPEG, blur, and noise settings. In contrast, changing the downsampling factor substantially affects classification for both interpolation modes. For bilinear factors $\times 3$ and $\times 5$, no test image is classified as bilinear: approximately 80% and 85% of inputs, respectively, are classified as JPEG. For bicubic factor $\times 3$, most inputs are classified as blur; at factor $\times 5$, all inputs are classified as blur. This motivates reporting Conditional UOT with both predicted and known family labels to separate classification errors from restoration performance.

D Experimental Details

D.1 Gaussian-to-Swiss-Roll Experiment

Distributions. We use five Gaussian source distributions $\mathbb{P}_k = \mathcal{N}(\mu_k, I_2)$ with means

$$\mu_1 = (0, 0), \quad \mu_2 = (3, 0), \quad \mu_3 = (0, 3), \quad \mu_4 = (-3, 0), \quad \mu_5 = (0, -3).$$

Method	FID ↓	LPIPS ↓	PSNR ↑
Bicubic $\times 5$ (training: $\times 4$)			
Degraded input	238.78	0.41126	20.67
Pooled UOT	**40.31**	**0.16003**	**20.09**
Conditional UOT + classifier	177.46	0.27675	<u>19.69</u>
SimNOT (ours)	<u>151.24</u>	<u>0.24771</u>	19.49
Conditional UOT + known family	21.89	0.09326	20.37
JPEG quality 20 (training: 25)			
Degraded input	36.78	0.05387	27.73
Pooled UOT	10.94	0.02463	28.52
Conditional UOT + classifier	**7.16**	<u>0.02094</u>	<u>28.89</u>
SimNOT (ours)	<u>7.30</u>	**0.02092**	**28.95**
Conditional UOT + known family	7.16	0.02094	28.89
Gaussian blur $\sigma = 1.75$ (training: 1.5)			
Degraded input	64.50	0.30943	24.13
Pooled UOT	<u>10.91</u>	**0.03433**	27.85
Conditional UOT + classifier	**9.88**	0.04010	**28.86**
SimNOT (ours)	11.82	<u>0.03488</u>	<u>28.38</u>
Conditional UOT + known family	9.88	0.04010	28.86
Gaussian noise $\sigma = 0.06$ (training: 0.05)			
Degraded input	65.16	0.06515	24.79
Pooled UOT	24.77	0.01584	30.65
Conditional UOT + classifier	<u>14.42</u>	<u>0.00932</u>	<u>32.18</u>
SimNOT (ours)	**11.26**	**0.00867**	**32.60**
Conditional UOT + known family	14.42	0.00932	32.18

Table 5: Generalization to stronger degradations. The evaluation protocol follows Table 4.

Degradation	Test parameter	Accuracy (%) ↑
Bicubic	$\times 3$	0.035
Bicubic	$\times 5$	0.000
Bilinear	$\times 3$	0.000
Bilinear	$\times 5$	0.000
JPEG	30	100.000
JPEG	20	100.000
Gaussian blur	1.25	99.995
Gaussian blur	1.75	100.000
Gaussian noise	0.04	100.000
Gaussian noise	0.06	100.000

Table 6: Degradation classification accuracy under parameter shifts. Each setting contains 20,259 test images. The classifier is trained only on the fixed degradation parameters specified in Appendix D.2.

To construct the target distribution, we first sample

$$\widetilde{y} = (t\cos t, t\sin t) + 0.75\,\epsilon, \qquad t \sim \mathrm{Unif}[1.5\pi, 4.5\pi], \qquad \epsilon \sim \mathcal{N}(0, I_2),$$

where t and ϵ are independent. We then standardize each coordinate using the empirical mean and standard deviation computed from a fixed calibration sample of $20{,}000$ points. These normalization statistics remain fixed throughout training and evaluation. Source and target samples are generated independently on demand.

Parametrization. The shared map $T_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ and the five potentials $v_{\omega_k} : \mathbb{R}^2 \to \mathbb{R}$ are fully connected networks with four hidden layers of width 256, ReLU activations, and linear output layers. The map is deterministic and uses no residual connection. All networks use Kaiming initialization with zero biases.

Training. We use the scaled quadratic transport cost $c(x, y) = \tau\|x - y\|_2^2$ with $\tau = 10^{-3}$ and the conjugate functions

$$\bar{\psi}(u) = \bar{\phi}(u) = 2\log(1 + \exp(u)) - 2\log 2.$$

Training consists of $100{,}000$ iterations, each comprising one update of the potentials followed by one update of the shared map. For the potential update, we independently sample 512 points from each source and a common batch of 512 target points. Fresh source batches of the same size are sampled for the map update. Contributions from the five sources are averaged with equal weights.

We regularize the potentials using an R1 penalty on target samples with coefficient $\gamma = 5$:

$$\frac{\gamma}{2K} \sum_{k=1}^{K} \mathbb{E}_{y\sim\mathbb{P}^*} \|\nabla_y v_{\omega_k}(y)\|_2^2, \qquad K = 5.$$

Visualization. We use fixed evaluation batches of $2{,}048$ points from each source and $2{,}048$ target points, sampled independently of the training batches. Figure 3 displays the outputs of the final map, applied directly to source samples without rejection sampling. The same target sample is shown in every panel.

D.2 CelebA Restoration Experiment

Data split. We use $202{,}592$ CelebA images at resolution 64×64, pooling the original dataset partitions and constructing a new split with random seed 0. We allocate $91{,}166$ images to the sources, $91{,}167$ to the clean target, and $20{,}259$ to the test set, corresponding approximately to a $45/45/10$ split. The source images are further divided into five disjoint subsets: $18{,}234$ images for bicubic downsampling and $18{,}233$ for each remaining degradation. Each image belongs to only one subset. The split is performed at the image level.

Only degraded source images and independently sampled clean target images are used for restoration training. The clean counterparts of source images are not provided as supervision. Paired degraded and clean test images are used exclusively for evaluation.

Degradations. We construct the five source distributions using the following transformations:

- *Bicubic downsampling.* Images are resized from 64×64 to 16×16 and back to 64×64, using bicubic interpolation at both stages.
- *Bilinear downsampling.* The same resizing procedure is applied using bilinear interpolation.
- *JPEG compression.* Images are encoded and decoded using JPEG quality 25.
- *Gaussian blur.* We apply a Gaussian filter with standard deviation 1.5 and kernel size 13×13.
- *Gaussian noise.* Independent Gaussian noise with standard deviation 0.05 is added to pixel values in $[0, 1]$, followed by clipping to $[0, 1]$.

Both resizing operations use `align_corners=False` without antialiasing. Network inputs are normalized to $[-1, 1]$. Noise realizations are sampled during training and fixed for evaluation.

Architectures. We parametrize the shared stochastic map $T_\theta(x, z)$ using an NCSN++ generator, with independent auxiliary noise $z \sim \mathcal{N}(0, I_{100})$. The generator uses a base channel width of 64, channel multipliers $(1, 1, 2, 2, 4, 4)$, two residual blocks per resolution, and attention at resolution 16×16. The noise embedding has dimension 256, and the output uses a hyperbolic tangent activation. The source index is not provided to the map.

Each of the five potentials is an independently parametrized `Discriminator_large` network with base width 64 and LeakyReLU activations with negative slope 0.2.

Training. We use the scaled quadratic cost $c(x, y) = \tau\|x - y\|_2^2$ with $\tau = 10^{-3}$, where the squared differences are summed over all pixels and channels. The conjugate functions are

$$\bar{\psi}(u) = \bar{\phi}(u) = 2\log(1 + \exp(u)) - 2\log 2.$$

We train for $100{,}000$ iterations, alternating one potential update with one map update. Each potential update uses 16 images from each source and a common batch of 16 clean target images. The map

update uses fresh batches of 16 images per source. Auxiliary noise is sampled independently for each input, and source contributions are averaged with equal weights.

We regularize the potentials using an R1 penalty on clean target samples with coefficient 5. Adam optimizers use $(\beta_1, \beta_2) = (0.5, 0.9)$, with initial learning rates of $2 \cdot 10^{-4}$ for the map and 10^{-4} for the potentials. Cosine learning-rate schedulers are advanced every 500 training iterations, with $T_{\max} = 700$ scheduler steps and $\eta_{\min} = 10^{-5}$. We initialize an exponential moving average of the map parameters at iteration 30,000 and subsequently update it with decay 0.999. All reported restoration results use the final EMA checkpoints at iteration 100,000.

Baseline implementations. Pooled UOT uses one stochastic map and one potential for the equally weighted mixture of the five sources. Conditional UOT uses one map and one potential, both conditioned on the degradation label. A learned degradation embedding is added to the generator's latent input and, for the potential, to the features after the first convolutional layer. The baselines use the same generator and potential backbones, source and target split, transport cost, and conjugate functions. They are trained for the same number of iterations, using 16 source images per degradation at each step.

Degradation classifier. For blind inference with Conditional UOT, we separately train a five-class classifier to predict the degradation family. We split the 91,166 source images into 82,049 training and 9,117 validation images using random seed 0. Each image is transformed using all five training degradations, yielding 410,245 training and 45,585 validation examples. The image split is performed before applying the degradations. The clean target subset and the 20,259 test images are not used to train or select the classifier.

The classifier is a compact ResNet-18-style network with stage widths $(32, 64, 128, 256)$, two residual blocks per stage, global average pooling, and dropout 0.1 before the final classification layer. It is trained from scratch for 100,000 steps using cross-entropy without label smoothing, AdamW with learning rate $3 \cdot 10^{-4}$ and weight decay 10^{-4}, and batch size 128. The learning rate follows a cosine schedule with a final value of $3 \cdot 10^{-6}$. We use random horizontal flips, gradient clipping at norm 5, and mixed-precision training. Degradation parameters remain fixed throughout training.

Validation is performed every 1,000 steps. We select the checkpoint with the highest validation accuracy, using the lowest validation loss to break ties. The selected checkpoint is from step 63,000. At inference, the predicted class is supplied to the conditional restoration model.

Evaluation on training degradations. Each of the 20,259 test images is randomly assigned one degradation using seed 0, with 4,052 images per degradation except Gaussian noise, which has 4,051. This assignment is shared across methods. We compute FID separately for the restored outputs of each degradation. Mean and maximum FID refer to the arithmetic mean and maximum of these five source-specific values. PSNR is evaluated against the corresponding clean reference images and averaged over all test images.

Evaluation under parameter shifts. We keep all trained models fixed and change one degradation parameter at a time. For each setting, the corresponding degradation is applied to all 20,259 test images. The evaluated parameters are

Degradation	Training	Evaluation
Bicubic factor	4	3, 5
Bilinear factor	4	3, 5
JPEG quality	25	30, 20
Blur standard deviation	1.5	1.25, 1.75
Noise standard deviation	0.05	0.04, 0.06

For downsampling factor s, the intermediate resolution is $\lfloor 64/s \rfloor \times \lfloor 64/s \rfloor$, followed by resizing back to 64×64 with the same interpolation mode. Thus, factors $\times 3$ and $\times 5$ correspond to intermediate resolutions 21×21 and 12×12, respectively.

We report FID, PSNR, and LPIPS. LPIPS is computed using the AlexNet backbone in TorchMetrics with inputs in $[-1, 1]$. For stochastic restoration, evaluation uses 100-dimensional Gaussian latent vectors with fixed random seed. Conditional UOT is evaluated both with predicted labels and with the known degradation family. The latter supplies only the family label, not the modified degradation parameter.

E LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

Scaling to many sources. SimNOT uses a shared transport map and a separate potential for each source distribution. While the transport network remains unchanged as the number of sources grows, maintaining and updating the potentials increases the training memory and computational requirements. Sharing parameters between potentials or sampling a subset of sources at each training step could improve scalability. These modifications would affect training only: inference uses a single shared map regardless of the number of sources.

Source grouping during training. Although SimNOT does not require paired source-target samples, it assumes that training samples are grouped by source distribution, so that each sample can be associated with the corresponding potential. This information allows each potential to evaluate the matching objective for its corresponding source. Such grouping is available in our image restoration setup, where the degradation process is known during training, but may be incomplete or unavailable in other applications. Extending SimNOT to partially observed source membership is a promising direction for future research. Once trained, the map can be applied directly to new inputs without specifying their source.