META OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the use of amortized optimization to predict optimal transport (OT) maps from the input measures, which we call *Meta OT*. This helps repeatedly solve similar OT problems between different measures by leveraging the knowledge and information present from past problems to rapidly predict and solve new problems. Otherwise, standard methods ignore the knowledge of the past solutions and suboptimally re-solve each problem from scratch. We instantiate Meta OT models in discrete and continuous (Wasserstein-2) settings between images, spherical data, and color palettes and use them to improve the computational time of standard OT solvers by multiple orders of magnitude.

1 INTRODUCTION

Optimal transportation (Villani, 2009; Ambrosio, 2003; Santambrogio, 2015; Peyré et al., 2019; Merigot and Thibert, 2021) is thriving in domains including economics (Galichon, 2016), reinforcement learning (Dadashi et al., 2021; Fickinger et al., 2021), style transfer (Kolkin et al., 2019), generative modeling (Arjovsky et al., 2017; Seguy et al., 2018; Huang et al., 2020; Rout et al., 2021), geometry (Solomon et al., 2015; Cohen et al., 2021), domain adaptation (Courty et al., 2017; Redko et al., 2019), signal processing (Kolouri et al., 2017), fairness (Jiang et al., 2020), and cell reprogramming (Schiebinger et al., 2019). A core component in these settings is to couple two measures (α, β) supported on domains $(\mathcal{X}, \mathcal{Y})$ by solving a transport optimization problem such as the *primal Kantorovich problem*, which is defined by:

$$\pi^\star(\alpha, \beta, c) \in \operatorname*{arg\,min}_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y), \tag{1}$$

where the *optimal coupling* $\pi^\star$ is a joint distribution over the product space, $\mathcal{U}(\alpha, \beta)$ is the set of admissible couplings between α and β, and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the *ground cost*, that represents a notion of distance between elements in $\mathcal{X}$ and elements in $\mathcal{Y}$.

Challenges. Unfortunately, solving eq. (1) *once* is computationally expensive between general measures and computationally cheaper alternatives are an active research topic: *Entropic optimal transport* (Cuturi, 2013) smooths the transport problem with an entropy penalty, and *sliced distances* (Kolouri et al., 2016; 2018; 2019; Deshpande et al., 2019) solve OT between 1-dimensional projections of the measures, where eq. (1) can be solved easily.

Furthermore, when an optimal transport method is deployed in practice, eq. (1) is not just solved a single time, but is *repeatedly* solved for new scenarios between different input measures (α, β). For example, the measures could be representations of images we care about optimally transporting between and in deployment we would receive a stream of new images to couple. Repeatedly solving optimal transport problems also comes up in the context of comparing seismic signals (Engquist and Froese, 2013) and in single-cell perturbations (Bunne et al., 2021; 2022b;a). Standard optimal transport solvers deployed in this setting would re-solve the optimization problems from scratch, but this ignores the shared structure and information present between different coupling problems.

Overview and outline. We study the use of amortized optimization and machine learning methods to rapidly solve multiple optimal transport problems and predict the solution from the input measures (α, β). This setting involves learning a *meta* model to predict the solution to the optimal transport problem, which we will refer to as *Meta Optimal Transport*. We learn Meta OT models to predict the solutions to optimal transport problems and significantly improve the computational time and number of iterations needed to solve eq. (1) between discrete (sect. 3.1) and continuous (sect. 3.2) measures. The paper is organized as follows: sect. 2 recalls the main concepts needed for the rest of the paper, in particular the formulations of the entropy regularized and unregularized optimal transport problems and the basic notions of amortized optimization; sect. 3 presents the Meta OT models and algorithms; and sect. 4 empirically demonstrates the effectiveness of Meta OT.

Settings that are not Meta OT. Meta OT is not useful in OT settings that do *not* involve *repeatedly* solving OT problems over a fixed distribution, including 1) standard generative modeling settings, such as Arjovsky et al. (2017) that estimate the OT distance between the data and model distributions, and 2) the out-of-sample setting of Seguy et al. (2018); Perrot et al. (2016) that couple measures and then extrapolate the map to larger measures containing the original measures.

2 Preliminaries and background

2.1 Dual optimal transport solvers

We review foundations of optimal transportation, following the notation of Peyré et al. (2019) in most places. The discrete setting often favors the entropic regularized version since it can be computed efficiently and in a parallelized way using the Sinkhorn algorithm. On the other hand, the continuous setting is often solved from samples using convex potentials. While the primal Kantorovich formulation in eq. (1) provides an intuitive problem description, optimal transport problems are rarely solved directly in this form due to the high-dimensionality of the couplings π and the difficulty of satisfying the coupling constraints $\mathcal{U}(\alpha, \beta)$. Instead, most computational OT solvers use the *dual* of eq. (1), which we build our Meta OT solvers on top of in discrete and continuous settings.

2.1.1 Entropic OT between discrete measures with the Sinkhorn algorithm

Let $\alpha := \sum_{i=1}^m a_i \delta_{x_i}$ and $\beta := \sum_{i=1}^n b_i \delta_{y_i}$ be *discrete* measures, where δ_z is a Dirac at point z and $a \in \Delta_{m-1}$ and $b \in \Delta_{n-1}$ are in the *probability simplex* defined by

$$\Delta_{k-1} := \{x \in \mathbb{R}^k : x \geq 0 \text{ and } \sum_i x_i = 1\}. \quad (2)$$

Algorithm 1 Sinkhorn($\alpha, \beta, c, \epsilon, f_0 = 0$)

```
for iteration i = 1 to N do
    $g_i \leftarrow \epsilon \log b - \epsilon \log \left(K^\top \exp\{f_{i-1}/\epsilon\}\right)$
    $f_i \leftarrow \epsilon \log a - \epsilon \log \left(K \exp\{g_i/\epsilon\}\right)$
end for
Compute $P_N$ from $f_N, g_N$ using eq. (6)
return $P_N \approx P^\star$
```

Discrete OT. In the discrete setting, eq. (1) simplifies to the *linear program*

$$P^\star(\alpha, \beta, c) \in \underset{P \in U(a,b)}{\arg\min} \langle C, P \rangle \qquad U(a,b) := \{P \in \mathbb{R}_+^{n \times m} : P1_m = a, \quad P^\top 1_n = b\} \quad (3)$$

where P is a *coupling matrix*, $P^\star(\alpha, \beta)$ is the *optimal* coupling, and the *cost* can be discretized as a matrix $C \in \mathbb{R}^{m \times n}$ with entries $C_{i,j} := c(x_i, y_j)$, and $\langle C, P \rangle := \sum_{i,j} C_{i,j} P_{i,j}$,

Entropic OT. The linear program above can be regularized adding the entropy of the coupling to smooth the objective as in Cominetti and Martín (1994); Cuturi (2013), resulting in:

$$P^\star(\alpha, \beta, c, \epsilon) \in \underset{P \in U(a,b)}{\arg\min} \langle C, P \rangle - \epsilon H(P) \quad (4)$$

where $H(P) := -\sum_{i,j} P_{i,j}(\log(P_{i,j}) - 1)$ is the discrete entropy of a coupling matrix P.

Entropic OT dual. As presented in Peyré et al. (2019, Prop. 4.4), the dual of eq. (4) is

$$f^\star, g^\star \in \underset{f \in \mathbb{R}^n, g \in \mathbb{R}^m}{\arg\max} \langle f, a \rangle + \langle g, b \rangle - \epsilon \left\langle \exp\{f/\epsilon\}, K \exp\{g/\epsilon\} \right\rangle, \quad K_{i,j} := \exp\{-C_{i,j}/\epsilon\}, \quad (5)$$

where $K \in \mathbb{R}^{m \times n}$ is the *Gibbs kernel* and the *dual variables* or *potentials* $f \in \mathbb{R}^n$ and $g \in \mathbb{R}^m$ are associated, respectively, with the marginal constraints $P1_m = a$ and $P^\top 1_n = b$. The optimal duals depend on the problem, e.g. $f^\star(\alpha, \beta, c, \epsilon)$, but we omit this dependence for notational simplicity.

Recovering the primal solution from the duals. Given optimal duals $f^\star$, $g^\star$ that solve eq. (5) the optimal coupling $P^\star$ to the primal problem in eq. (4) can be obtained by

$$P^\star_{i,j}(\alpha, \beta, c, \epsilon) := \exp\{f^\star_i/\epsilon\} K_{i,j} \exp\{g^\star_j/\epsilon\} \qquad (K \text{ is defined in eq. (5)}) \quad (6)$$

The Sinkhorn algorithm. Algorithm 1 summarizes the log-space version, which takes closed-form block coordinate ascent updates on eq. (5) obtained from the first-order optimality conditions (Peyré et al., 2019, Remark 4.21). We will use it to fine-tune predictions made by our Meta OT models.

Computing the error. Standard implementations of the Sinkhorn algorithm, such as Flamary et al. (2021); Cuturi et al. (2022), measure the error of a candidate dual solution (f, g) by computing the deviation from the marginal constraints, which we will also use in comparing our solution quality:

$$\mathrm{err}(f, g; \alpha, \beta, c) := \|P1_m - a\|_1 + \|P^\top 1_n - b\|_1 \qquad \text{(compute } P \text{ from eq. (6))} \tag{7}$$

Mapping between the duals. The first-order optimality conditions of eq. (5) also provide an equivalence between the optimal dual potentials that we will make use of:

$$g(f; b, c) := \epsilon \log b - \epsilon \log\left(K^\top \exp\{f/\epsilon\}\right). \tag{8}$$

2.1.2 Dual Wasserstein-2 OT between continuous (Euclidean) measures

Let α and β be continuous measures in Euclidean space $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ (with α absolutely continuous with respect to the Lebesgue measure) and the ground cost be the squared Euclidean distance $c(x, y) := \|x - y\|_2^2$. Then the minimum of eq. (1) defines the square of the *Wasserstein-2* distance:

```
Algorithm 2 W2GN($\alpha, \beta, \varphi_0$)
for iteration $i$ = 1 to $N$ do
    Sample from $(\alpha, \beta)$ and estimate $\mathcal{L}(\varphi_{i-1})$
    Update $\varphi_i$ with approximation to $\nabla_\varphi \mathcal{L}(\varphi_{i-1})$
end for
return $T_N(\cdot) := \nabla_x \psi_{\varphi_N}(\cdot) \approx T^\star(\cdot)$
```

$$W_2^2(\alpha, \beta) := \min_{\pi \in \mathcal{U}(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{Y}} \|x - y\|_2^2 \mathrm{d}\pi(x, y) = \min_T \int_{\mathcal{X}} \|x - T(x)\|_2^2 \mathrm{d}\alpha(x), \tag{9}$$

where T is a *transport map* pushing α to β, i.e. $T_\# \alpha = \beta$ with the *pushforward operator* defined by $T_\# \alpha(B) := \alpha(T^{-1}(B))$ for any measurable set B.

Convex dual potentials. The primal form in eq. (9) is difficult to solve, as in the discrete setting, due to the difficulty of representing the coupling and satisfying the constraints. Makkuva et al. (2020); Taghvaei and Jalali (2019); Korotin et al. (2019; 2021b; 2022) propose to instead solve the dual:

$$\psi^\star(\,\cdot\,; \alpha, \beta) \in \operatorname*{arg\,min}_{\psi \in \text{convex}} \int_{\mathcal{X}} \psi(x) \mathrm{d}\alpha(x) + \int_{\mathcal{Y}} \overline{\psi}(y) \mathrm{d}\beta(y), \tag{10}$$

where ψ is a convex function referred to as a *convex potential*, and $\overline{\psi}(y) := \max_{x \in \mathcal{X}} \langle x, y \rangle - \psi(x)$ is the *Legendre-Fenchel transform* or *convex conjugate* of ψ (Fenchel, 1949; Rockafellar, 2015). The potential ψ is often approximated with an input-convex neural network (ICNN) (Amos et al., 2017).

Recovering the primal solution from the dual. Given an optimal dual $\psi^\star$ for eq. (10), Brenier (1991) remarkably shows that an optimal map $T^\star$ for eq. (9) can be obtained with differentiation:

$$T^\star(x) = \nabla_x \psi^\star(x). \tag{11}$$

Wasserstein-2 Generative Networks (W2GNs). Korotin et al. (2019) model ψ_φ and $\overline{\psi_\varphi}$ in eq. (10) with two separate ICNNs parameterized by φ. The separate model for $\overline{\psi_\varphi}$ is useful because the conjugate operation in eq. (10) becomes computationally expensive. They optimize the loss:

$$\mathcal{L}(\varphi) := \underbrace{\mathop{\mathbb{E}}_{x \sim \alpha} [\psi_\varphi(x)] + \mathop{\mathbb{E}}_{y \sim \beta} \left[\langle \nabla \overline{\psi_{\cancel{\varphi}}}(y), y \rangle - \psi_\varphi(\nabla \overline{\psi_{\cancel{\varphi}}}(y))\right]}_{\text{Cyclic monotone correlations (dual objective)}} + \gamma \underbrace{\mathop{\mathbb{E}}_{y \sim \beta} \|\nabla \psi_{\cancel{\varphi}} \circ \nabla \overline{\psi_\varphi}(y) - y\|_2^2}_{\text{Cycle-consistency regularizer}}, \tag{12}$$

where $\cancel{\varphi}$ is a detached copy of the parameters and γ is a hyper-parameter. The first term are the *cyclic monotone correlations* (Chartrand et al., 2009; Taghvaei and Jalali, 2019), that optimize the dual objective in eq. (10), and the second term provides *cycle consistency* (Zhu et al., 2017) to estimate the conjugate $\overline{\psi}$. Algorithm 2 shows how $\mathcal{L}$ is typically optimized using samples from the measures, which we use to fine-tune Meta OT predictions.

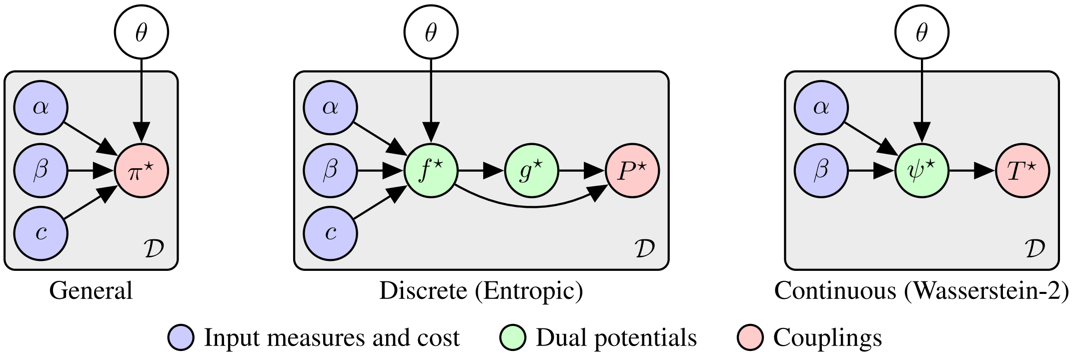

Figure 1: Meta OT uses objective-based amortization for optimal transport. In the general formulation, the *parameters* θ capture shared structure in the *optimal couplings* $\pi^\star$ between multiple input measures and costs over some *distribution* $\mathcal{D}$. In practice, we learn this shared structure over the *dual potentials* which map back to the coupling: $f^\star$ in discrete settings and $\psi^\star$ in continuous ones.

2.2 Amortized optimization and learning to optimize

Our paper is an application of amortized optimization methods that predict the solutions of optimization problems, as surveyed in, e.g., Chen et al. (2021); Amos (2022). We use the basic setup from Amos (2022), which considers unconstrained continuous optimization problems of the form

$$z^\star(\phi) \in \operatorname*{arg\,min}_z J(z;\phi), \tag{13}$$

where J is the objective, $z \in \mathcal{Z}$ is the *domain*, and $\phi \in \Phi$ is some *context* or *parameterization*. In other words, the context conditions the objective but is not optimized over. Given a *distribution over contexts* $\mathcal{P}(\phi)$, we learn a model $\hat{z}_\theta$ parameterized by θ to approximate eq. (13), i.e. $\hat{z}_\theta(\phi) \approx z^\star(\phi)$. J will be differentiable for us, so we optimize the parameters using *objective-based learning* with

$$\min_\theta \mathop{\mathbb{E}}_{\phi\sim\mathcal{P}(\phi)} J(\hat{z}_\theta(\phi);\phi), \tag{14}$$

which does *not* require ground-truth solutions $z^\star$ and can be optimized with a gradient-based solver. While we focus on optimizing eq. (14) because we do not assume easy access to ground-truth solutions $z^\star(\phi)$, one alternative is *regression-based learning* if the solutions are easily available:

$$\min_\theta \mathop{\mathbb{E}}_{\phi\sim\mathcal{P}(\phi)} \|z^\star(\phi) - \hat{z}_\theta(\phi)\|_2^2. \tag{15}$$

3 Meta Optimal Transport

Figure 1 illustrates our key contribution of connecting objective-based amortization in eq. (14) to optimal transport. We consider solving *multiple* OT problems and learning shared structure and correlations between them. We denote a joint *meta-distribution* over the input measures and costs with $\mathcal{D}(\alpha,\beta,c)$, which we call *meta* to distinguish it from the measures α,β.

In general, we could introduce a model that directly predicts the primal solution to eq. (1), i.e. $\pi_\theta(\alpha,\beta,c) \approx \pi^\star(\alpha,\beta,c)$ for $(\alpha,\beta,c)\sim\mathcal{D}$. This is difficult for the same reason why most computational methods do not operate directly in the primal space: the optimal coupling is often a high-dimensional joint distribution with non-trivial marginal constraints. We instead turn to predicting the dual variables used by today's solvers.

3.1 Meta OT between discrete measures

We build on standard methods for entropic OT reviewed in sect. 2.1.1 between discrete measures $\alpha := \sum_{i=1}^m a_i\delta_{x_i}$ and $\beta := \sum_{i=1}^n b_i\delta_{x_i}$ with $a \in \Delta_{m-1}$ and $b \in \Delta_{n-1}$ coupled using a cost c. In the Meta OT setting, the measures and cost are the contexts for amortization and sampled from a *meta-distribution*, i.e. $(\alpha,\beta,c)\sim\mathcal{D}(\alpha,\beta,c)$. For example, sects. 4.1 and 4.2 considers meta-distributions over the weights of the atoms, i.e. $(a,b)\sim\mathcal{D}$, where $\mathcal{D}$ is a distribution over $\Delta_{m-1}\times\Delta_{n-1}$.

Algorithm 3 Training Meta OT

```
Initialize amortization model with $\theta_0$
for iteration $i$ do
    Sample $(\alpha, \beta, c) \sim \mathcal{D}$
    Predict duals $\hat{f}_\theta$ or $\hat{\varphi}_\theta$ on the sample
    Estimate the loss in eq. (17) or eq. (18)
    Update $\theta_{i+1}$ with a gradient step
end for
```

Algorithm 4 Fine-tuning with Sinkhorn

```
Predict duals $\hat{f}_\theta(\alpha, \beta, c)$
return Sinkhorn($\alpha, \beta, c, \epsilon, \hat{f}_\theta$)
```

Algorithm 5 Fine-tuning with W2GN

```
Predict dual ICNN parameters $\hat{\varphi}_\theta(\alpha, \beta, c)$
return W2GN($\alpha, \beta, c, T, \hat{\varphi}_\theta$)
```

Amortization objective. We will seek to predict the *optimal* potential. At optimality, the pair of potentials are related to each other via eq. (8), i.e. $g(f; \alpha, \beta, c) := \epsilon \log b - \epsilon \log \left(K^\top \exp\{f/\epsilon\} \right)$ where $K \in \mathbb{R}^{m \times n}$ is the *Gibbs kernel* from eq. (5). Hence, it is sufficient to predict one of the potentials, e.g. f, and recover the other. We thus re-formulate eq. (5) to just optimize over f with

$$f^\star(\alpha, \beta, c, \epsilon) \in \operatorname*{arg\,min}_{f \in \mathbb{R}^n} J(f; \alpha, \beta, c), \tag{16}$$

where $-J(f; \alpha, \beta, c) := \langle f, a \rangle + \langle g, b \rangle - \epsilon \left\langle \exp\{f/\epsilon\}, K \exp\{g/\epsilon\} \right\rangle$ is the (negated) dual objective. Even though most solvers optimize over f and g jointly as in eq. (16), amortizing over these would likely need: 1) to have a higher capacity than a model just predicting f, and 2) to learn how f and g are connected through eq. (8) while in eq. (16) we explicitly provide this knowledge.

Amortization model. We predict the solution to eq. (16) with $\hat{f}_\theta(\alpha, \beta, c)$ parameterized by θ, resulting in a computationally efficient approximation $\hat{f}_\theta \approx f^\star$. Here we use the notation $\hat{f}_\theta(\alpha, \beta, c)$ to mean that the model $\hat{f}_\theta$ depends on *representations* of the input measures and cost. In our settings, we define $\hat{f}_\theta$ as a fully-connected MLP mapping from the atoms of the measures to the duals.

Amortization loss. Applying objective-based amortization from eq. (14) to the dual in eq. (16) completes our learning setup. Our model should best-optimize the expectation of the dual objective

$$\min_\theta \mathop{\mathbb{E}}_{(\alpha, \beta, c) \sim \mathcal{D}} J(\hat{f}_\theta(\alpha, \beta, c); \alpha, \beta, c), \tag{17}$$

which is appealing as it does not require ground-truth solutions $f^\star$. Algorithm 3 shows a basic training loop for eq. (17) using a gradient-based optimizer such as Adam (Kingma and Ba, 2014).

Sinkhorn fine-tuning. The dual prediction made by $\hat{f}_\theta$ with an associated $\hat{g}$ can easily be input as the initialization to a standard Sinkhorn solver as shown in algorithm 4. This allows us to deploy the predicted potential with Sinkhorn to obtain the optimal potentials with only a few extra iterations.

On accelerated solvers. Here we have only considered fine-tuning the Meta OT prediction with a log-Sinkhorn solver. Meta OT can also be combined with accelerated variants of entropic OT solvers such as Thibault et al. (2017); Altschuler et al. (2017); Alaya et al. (2019); Lin et al. (2019) that would otherwise solve every problem from scratch.

3.2 META OT BETWEEN CONTINUOUS MEASURES (WASSERSTEIN-2)

We take an analogous approach to predicting the Wasserstein-2 map between continuous measures for Wasserstein-2 as reviewed in sect. 2.1.2. Here the measures α, β are supported in continuous space $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and we focus on computing Wasserstein-2 couplings from instances sampled from a *meta-distribution* $(\alpha, \beta) \sim \mathcal{D}(\alpha, \beta)$. The cost c is not included in $\mathcal{D}$ as it remains fixed to the squared Euclidean cost everywhere here.

One challenge here is that the optimal dual potential $\psi^\star(\,\cdot\,; \alpha, \beta)$ in eq. (10) is a convex function and not simply a finite-dimensional real vector. The dual potentials in this setting are approximated by, e.g., an ICNN. We thus propose a *Meta ICNN* that predicts the *parameters* φ of an ICNN ψ_φ that approximates the optimal dual potentials, which can be seen as a hypernetwork (Stanley et al., 2009; Ha et al., 2016). The dual prediction made by $\hat{\varphi}_\theta$ can easily be input as the initial value to a standard W2GN solver as shown in algorithm 5. App. B discusses other modeling choices we considered: we tried models based on MAML (Finn et al., 2017) and neural processes (Garnelo et al., 2018b;a).

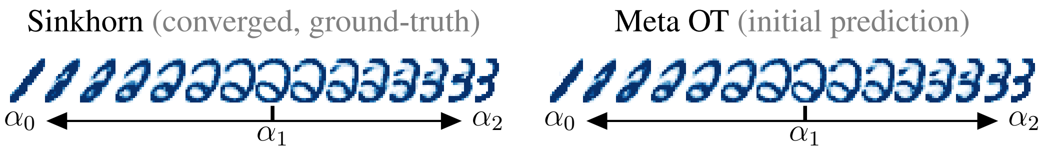
Sinkhorn (converged, ground-truth)
Meta OT (initial prediction)
α_0
α_1
α_2
α_0
α_1
α_2

Figure 2: Interpolations between MNIST test digits using couplings obtained from (left) solving the problem with Sinkhorn, and (right) Meta OT model's initial prediction, which is ≈**100** times computationally cheaper and produces a nearly identical coupling.

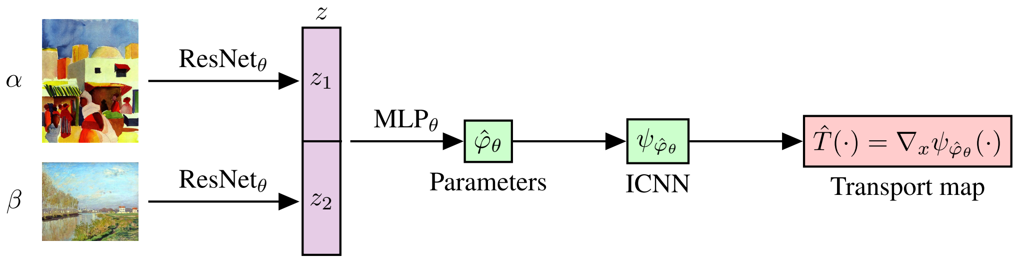

Figure 3: A Meta ICNN for image-based input measures. A shared ResNet processes the input measures α and β into latents z that are decoded with an MLP into the parameters φ of an ICNN dual potential ψ_φ. The derivative of the ICNN provides the transport map $\hat{T}$.

Table 1: Sinkhorn runtime (seconds) to reach a marginal error of 10^{-2}. Meta OT's initial prediction takes $\approx 5 \cdot 10^{-5}$ seconds.

Initialization	MNIST	Spherical
Zeros (t_{zeros})	$4.5 \cdot 10^{-3} \pm 1.5 \cdot 10^{-3}$	0.88 ± 0.13
Gaussian	$4.1 \cdot 10^{-3} \pm 1.2 \cdot 10^{-3}$	$0.56 \pm 9.9 \cdot 10^{-2}$
Meta OT (t_{Meta})	$2.3 \cdot 10^{-3} \pm 9.2 \cdot 10^{-6}$	$7.8 \cdot 10^{-2} \pm 3.4 \cdot 10^{-2}$
Improvement ($t_{\text{zeros}}/t_{\text{Meta}}$)	1.96	11.3

Table 2: Color transfer runtimes and values.

	Iter	Runtime (s)	Dual Value
Meta OT + W2GN	None	$3.5 \cdot 10^{-3} \pm 2.7 \cdot 10^{-4}$	$0.90 \pm 6.08 \cdot 10^{-2}$
	1k	$0.93 \pm 2.27 \cdot 10^{-2}$	$1.0 \pm 2.57 \cdot 10^{-3}$
	2k	$1.84 \pm 3.78 \cdot 10^{-2}$	$1.0 \pm 5.30 \cdot 10^{-3}$
W2GN	1k	$0.90 \pm 1.62 \cdot 10^{-2}$	$0.96 \pm 2.62 \cdot 10^{-2}$
	2k	$1.81 \pm 3.05 \cdot 10^{-2}$	$0.99 \pm 1.14 \cdot 10^{-2}$

We report the mean and standard deviation across 10 test instances.

Amortization objective. We build on the W2GN formulation (Korotin et al., 2019) and seek parameters $\varphi^\star$ optimizing the dual ICNN potentials ψ_φ and $\overline{\psi_\varphi}$ with $\mathcal{L}(\varphi; \alpha, \beta)$ from eq. (12). We chose W2GN due to the stability, but could also easily use other losses optimizing ICNN potentials.

Amortization model: the Meta ICNN. We predict the solution to eq. (12) with $\hat{\varphi}_\theta(\alpha, \beta)$ parameterized by θ, resulting in a computationally efficient approximation to the optimum $\hat{\varphi}_\theta \approx \varphi^\star$. Figure 3 instantiates a convolutional Meta ICNN model using a ResNet-18 (He et al., 2016) architecture for coupling image-based measures. We again emphasize that α, β used with the model here are *representations* of measures, which in our cases are simply images.

Amortization loss. Applying objective-based amortization from eq. (14) to the W2GN loss in eq. (12) completes our learning setup. Our model should best-optimize the expectation of the loss:

$$\min_\theta \underset{(\alpha,\beta)\sim\mathcal{D}}{\mathbb{E}} \mathcal{L}(\hat{\varphi}_\theta(\alpha, \beta); \alpha, \beta). \tag{18}$$

As in the discrete setting, it does not require ground-truth solutions $\varphi^\star$ and we learn it with Adam.

4 Experiments

We demonstrate how Meta OT models improve the convergence of the state-of-the-art solvers in settings where solving multiple OT problems naturally arises. We implemented our code in JAX (Bradbury et al., 2018) as an extension to the the Optimal Transport Tools (OTT) package (Cuturi et al., 2022). App. C covers further experimental and implementation details, and shows that all of our experiments take a few hours to run on our single Quadro GP100 GPU. We will open source the code to reproduce all of our experiments.

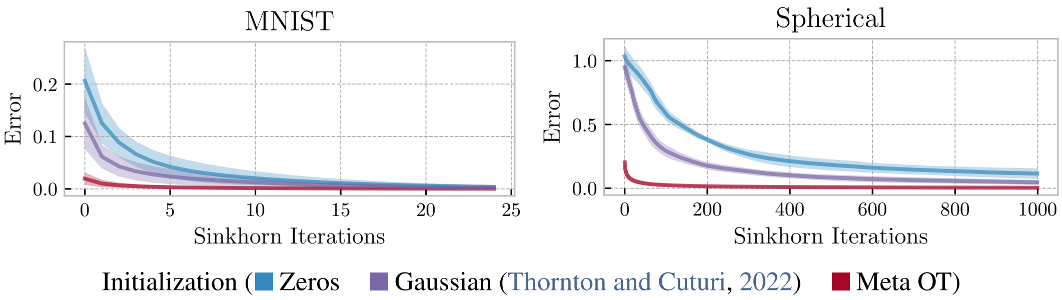

Figure 4: Meta OT successfully predicts warm-start initializations that significantly improve the convergence of Sinkhorn iterations on test data. The error is the marginal error defined in eq. (7).

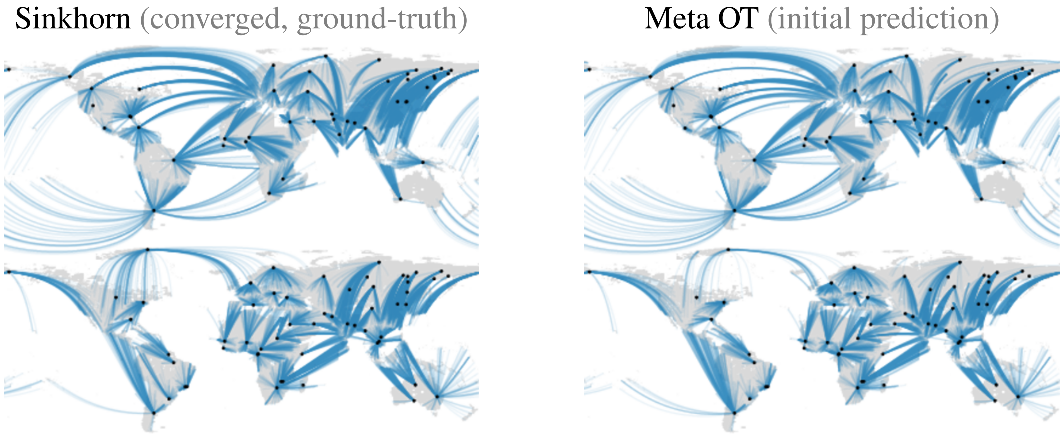
Sinkhorn (converged, ground-truth)
Meta OT (initial prediction)

Figure 5: Test set coupling predictions of the spherical transport problem. Meta OT's initial prediction is $\approx\mathbf{37500}$ times faster than solving Sinkhorn to optimality. Supply locations are shown as black dots and the blue lines show the spherical transport maps T going to demand locations at the end. The sphere is visualized with the Mercator projection.

4.1 Discrete OT between MNIST digits

Images provide a natural setting for Meta OT where the distribution over images provide the meta-distribution $\mathcal{D}$ over OT problems. Given a pair of images α_0 and α_1, each grayscale image is cast as a discrete measure in 2-dimensional space where the intensities define the probabilities of the atoms. The goal is to compute the optimal transport interpolation between the two measures as in, e.g., Peyré et al. (2019, §7). Formally, this means computing the optimal coupling $P^\star$ by solving the entropic optimal transport problem between α_0 and α_1 and computing the interpolates as $\alpha_t = (t\,\mathrm{proj}_y + (1-t)\,\mathrm{proj}_x)_{\#}P^\star$, for $t \in [0,1]$, where $\mathrm{proj}_x(x,y) := x$ and $\mathrm{proj}_y(x,y) = y$. We selected $\epsilon = 10^{-2}$ as app. A shows that it gives interpolations that are not too blurry or sharp.

Our Meta OT model $\hat{f}_\theta$ (sect. 3.1) is an MLP that predicts the transport map between pairs of MNIST digits. We train on every pair from the standard training dataset. Figure 2 shows that even without fine-tuning, Meta OT's predicted Wasserstein interpolations between the measures are close to the ground-truth interpolations obtained from running the Sinkhorn algorithm to convergence. We then fine-tune Meta OT's prediction with Sinkhorn as in algorithm 4. Figure 4 shows that the near-optimal predictions can be quickly refined in fewer iterations than running Sinkhorn with the default initialization, and table 1 shows the runtime required to reach an error threshold of 10^{-2}, showing that the Meta OT initialization help solve the problems faster by an order of magnitude. We compare our learned initialization to the standard zero initialization, as well as the Gaussian initialization proposed in Thornton and Cuturi (2022), which takes a continuous Gaussian approximation of the measures and initializes the potentials to be the known coupling between the Gaussians. This Gaussian initialization assumes the squared Euclidean cost, which is not the case in our spherical transport problem, but we find it is still helpful over the zero initialization.

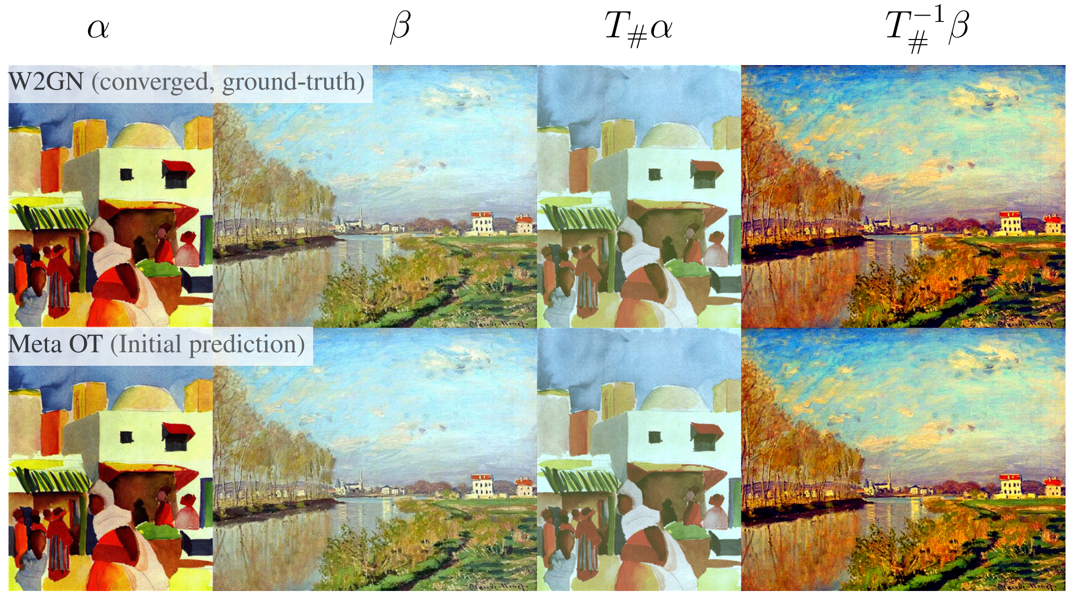

Figure 6: Color transfers with a Meta ICNN on test pairs of images. The objective is to optimally transport the continuous RGB measure of the first image α to the second β, producing an invertible transport map T. Meta OT's prediction is $\approx$**1000** times faster than training W2GN from scratch. The image generating α is Market in Algiers by August Macke (1914) and β is Argenteuil, The Seine by Claude Monet (1872), obtained from WikiArt.

4.2 Discrete OT for supply-demand transportation on spherical data

We next set up a synthetic transport problem between supply and demand locations where the supply and demands may change locations or quantities frequently, creating another Meta OT setting to be able to rapidly solve the new instances. We specifically consider measures living on the 2-sphere defined by $\mathcal{S}_2 := \{x \in \mathbb{R}^3 : \|x\| = 1\}$, i.e. $\mathcal{X} = \mathcal{Y} = \mathcal{S}_2$, with the transport cost given by the spherical distance $c(x, y) = \arccos(\langle x, y \rangle)$. We then randomly sample supply locations uniformly from Earth's landmass and demand locations from Earth's population density to induce a class of transport problems on the sphere obtained from the CC-licensed dataset from Doxsey-Whitfield et al. (2015). Figure 5 shows that the predicted transport maps on test instances are close to the optimal maps obtained from Sinkhorn to convergence. Similar to the MNIST setting, fig. 4 and table 1 show improved convergence and runtime.

4.3 Continuous Wasserstein-2 color transfer

The problem of color transfer between two images consists in mapping the color palette of one image into the other one. The images are required to have the same number of channels, for example RGB images. The continuous formulation that we use from Korotin et al. (2019), takes i.e. $\mathcal{X} = \mathcal{Y} = [0, 1]^3$ with c being the squared Euclidean distance. We collected $\approx$200 public domain images from WikiArt and trained a Meta ICNN model from sect. 3.2 to predict the color transfer maps between every pair of them. Figure 6 shows the predictions on test pairs and fig. 7 shows the convergence in comparison to the standard W2GN learning. Table 2 reports runtimes and app. E shows additional results.

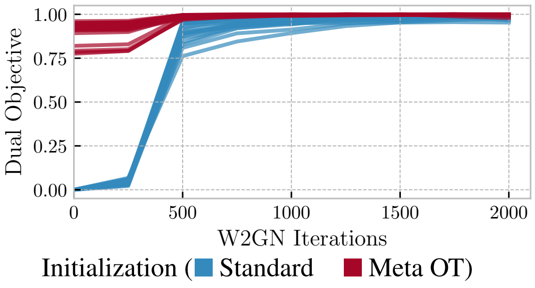

Figure 7: Convergence on color transfer test instances using W2GN. Meta ICNNs predicts warm-start initializations that significantly improve the (normalized) dual objective values.

5 Related work

Efficiently estimating OT maps. To compute OT maps with fixed cost between pairs of measures efficiently, neural OT models (Korotin et al., 2019; 2021a; Mokrov et al., 2021; Korotin et al., 2021b) leverage ICNNs to estimate maps between continuous high-dimensional measures given samples from these, and Litvinenko et al. (2021); Scetbon et al. (2021); Forrow et al. (2019); Sommerfeld et al. (2019); Scetbon et al. (2022); Muzellec and Cuturi (2019); Bonet et al. (2021) leverage structural assumptions on coupling and cost matrices to reduce the computational and memory complexity. In the meta-OT setting, we consider learning to rapidly compute OT mappings between new pairs measures. All these works can hence potentially benefit from an acceleration effect by leveraging amortization similarly.

Embedding measures where OT distances are discriminative. Effort has been invested in learning encodings/projections of measures through a nested optimization problem, which aims to find discriminative embeddings of the measures to be compared (Genevay et al., 2018; Deshpande et al., 2019; Nguyen and Ho, 2022). While these works share an encoder and/or a projection across task with the aim of leveraging more discriminative alignments (and hence an OT distance with a metric different from the Euclidean metric), our work differs in the sense that we find good initializations to solve the OT problem itself with fixed cost more efficiently across tasks.

Optimal transport and amortization. Few previous works in the OT literature leverage amortization. Courty et al. (2018) learn a latent space in which the Wasserstein distance between the measure's embeddings is equivalent to the Euclidean distance. Concurrent work (Nguyen and Ho, 2022) amortizes the estimation of the optimal projection in the max-sliced objective, which differs from our work where we instead amortize the estimation of the optimal coupling directly. Also, Lacombe et al. (2021) learns to predict Wasserstein barycenters of pixel images by training a convolutional networks that, given images as input, outputs their barycenters. Our work is hence a generalization of this pixel-based work to general measures – both discrete and continuous. A limitation of Lacombe et al. (2021) is that it does not provide alignments, as the amortization networks predicts the barycenter directly rather than individual couplings.

6 Conclusions, future directions, and limitations

We have presented foundations for modeling and learning to solve OT problems with Meta OT by using amortized optimization to predict optimal transport plans. This works best in applications that require solving multiple OT problems with shared structure. We instantiated it to speed up entropic regularized optimal transport and unregularized optimal transport with squared cost by multiple orders of magnitude. We envision extensions of the work in:

1. **Meta OT models**. While we mostly consider models based on hypernetworks, other meta-learning paradigms can be connected in. In the discrete setting, we only considered settings where the cost remains fixed, but the Meta OT model can also be conditioned on the cost by considering the entire cost matrix as an input (which may be too large for most models to handle), or considering a lower-dimensional parameterization of the cost that changes between the Meta OT problem instances.

2. **OT algorithms**. While we instantiated models on top of log-Sinkhorn and W2GN, Meta OT could be built on top of other methods.

3. **OT applications** that are computationally expensive and repeatedly solved, e.g. in multi-marginal and barycentric settings, or for Gromov-Wasserstein distances between metric-measure spaces.

Limitations. While we have illustrated successful applications of Meta OT, it is also important to understand the limitations: 1) **Meta OT does not make previously intractable problems tractable.** All of the baseline OT solvers we consider solve our problems within milliseconds or seconds. 2) **Out-of-distribution generalization.** Meta OT may not generate good predictions on instances that are not close to the training OT problems from the meta-distribution $\mathcal{D}$ over the measures and cost. If the model makes a bad prediction, one fallback option is to re-solve the instance from scratch.

Reproducibility statement

We have tried to clearly articulate all of the relevant details in the text so that this paper can be completely reproduced from the contents of this paper alone. We will also open-source the full source code behind every experimental result, table, and figure in this paper, and will anonymously send it with this submission.

References

Mokhtar Z Alaya, Maxime Berar, Gilles Gasso, and Alain Rakotomamonjy. Screening sinkhorn algorithm for regularized optimal transport. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Luigi Ambrosio. Lecture notes on optimal transport problems. In *Mathematical aspects of evolving interfaces*, pages 1–52. Springer, 2003.

Brandon Amos. Tutorial on amortized optimization for learning to optimize over continuous domains. *arXiv preprint arXiv:2202.00665*, 2022.

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

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

Clément Bonet, Titouan Vayer, Nicolas Courty, François Septier, and Lucas Drumetz. Subspace detours meet gromov–wasserstein. *Algorithms*, 14(12):366, 2021.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. *GitHub*, 2018. URL http://github.com/google/jax.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Ratsch. Learning single-cell perturbation responses using neural optimal transport. *bioRxiv*, 2021.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional monge maps. *arXiv preprint arXiv:2206.14262*, 2022a.

Charlotte Bunne, Laetitia Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal optimal transport modeling of population dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR, 2022b.

Rick Chartrand, Brendt Wohlberg, Kevin Vixie, and Erik Bollt. A gradient descent solution to the monge-kantorovich problem. *Applied Mathematical Sciences*, 3(22):1071–1080, 2009.

Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to optimize: A primer and a benchmark. *arXiv preprint arXiv:2103.12828*, 2021.

Samuel Cohen, Brandon Amos, and Yaron Lipman. Riemannian convex potential maps. In *International Conference on Machine Learning*, pages 2028–2038. PMLR, 2021.

Roberto Cominetti and J San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1):169–187, 1994.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017.

Nicolas Courty, Rémi Flamary, and Mélanie Ducoffe. Learning wasserstein embeddings. In *International Conference on Learning Representations*, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

Robert Dadashi, Léonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal wasserstein imitation learning. In *ICLR 2021-Ninth International Conference on Learning Representations*, 2021.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.

Erin Doxsey-Whitfield, Kytt MacManus, Susana B Adamo, Linda Pistolesi, John Squires, Olena Borkovska, and Sandra R Baptista. Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4. *Papers in Applied Geography*, 1(3):226–234, 2015.

Bjorn Engquist and Brittany D Froese. Application of the wasserstein metric to seismic signals. *arXiv preprint arXiv:1311.4581*, 2013.

Werner Fenchel. On conjugate convex functions. *Canadian Journal of Mathematics*, 1(1):73–77, 1949.

Arnaud Fickinger, Samuel Cohen, Stuart Russell, and Brandon Amos. Cross-domain imitation learning via optimal transport. In *International Conference on Learning Representations*, 2021.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/finn17a.html.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2454–2465. PMLR, 2019.

Alfred Galichon. Optimal transport methods in economics. In *Optimal Transport Methods in Economics*. Princeton University Press, 2016.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2018a.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/genevay18a.html.

David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

Chin-Wei Huang, Ricky TQ Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. In *International Conference on Learning Representations*, 2020.

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

John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90, 2007.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pages 862–872. PMLR, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019.

Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.

Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo K Rohde. Generalized sliced wasserstein distances. *arXiv preprint arXiv:1902.00434*, 2019.

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. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022.

Julien Lacombe, Julie Digne, Nicolas Courty, and Nicolas Bonneel. Learning to generate wasserstein barycenters, 2021. URL https://openreview.net/forum?id=2ioNazs6lvw.

Lingxiao Li, Aude Genevay, Mikhail Yurochkin, and Justin Solomon. Continuous regularized wasserstein barycenters. *arXiv preprint arXiv:2008.12534*, 2020.

Tianyi Lin, Nhat Ho, and Michael I Jordan. On the acceleration of the sinkhorn and greenkhorn algorithms for optimal transport. *arXiv preprint arXiv:1906.01437*, 2019.

Alexander Litvinenko, Youssef Marzouk, Hermann G Matthies, Marco Scavino, and Alessio Spantini. Computing f-divergences and distances of high-dimensional probability density functions–low-rank tensor approximations. *arXiv preprint arXiv:2111.07164*, 2021.

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

Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

Quentin Merigot and Boris Thibert. Optimal transport: discretization and algorithms. In *Handbook of Numerical Analysis*, volume 22, pages 133–212. Elsevier, 2021.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 34:15243–15256, 2021.

Boris Muzellec and Marco Cuturi. Subspace detours: Building transport plans that are optimal on subspace projections. *Advances in Neural Information Processing Systems*, 32, 2019.

Khai Nguyen and Nhat Ho. Amortized projection optimization for sliced wasserstein generative models. *arXiv preprint arXiv:2203.13417*, 2022.

Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

Travis E Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20, 2007.

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. *Advances in Neural Information Processing Systems*, 29, 2016.

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.

Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 849–858. PMLR, 2019.

Ralph Tyrell Rockafellar. Convex analysis. In *Convex analysis*. Princeton university press, 2015.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*, 2021.

Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations*, 2018.

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

Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-rank sinkhorn factorization. In *International Conference on Machine Learning*, pages 9344–9354. PMLR, 2021.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time gromov wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pages 19347–19365. PMLR, 2022.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4): 928–943.e22, 2019. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2019.01.006. URL https://www.sciencedirect.com/science/article/pii/S009286741930039X.

Vivien Seguy, Bharath Bhushan Damodaran, Remi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large scale optimal transport and mapping estimation. In *International Conference on Learning Representations*, 2018.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

Max Sommerfeld, Jörn Schrieber, Yoav Zemel, and Axel Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *J. Mach. Learn. Res.*, 20:105–1, 2019.

Kenneth O Stanley, David B D'Ambrosio, and Jason Gauci. A hypercube-based encoding for evolving large-scale neural networks. *Artificial life*, 15(2):185–212, 2009.

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.

Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2021.

Alexis Thibault, Lenaic Chizat, Charles Dossal, and Nicolas Papadakis. Overrelaxed sinkhorn-knopp algorithm for regularized optimal transport. *arXiv preprint arXiv:1711.01851*, 2017.

James Thornton and Marco Cuturi. Rethinking initialization of the sinkhorn algorithm. *arXiv preprint arXiv:2206.07630*, 2022.

Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22, 2011.

Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Omry Yadan. Hydra - a framework for elegantly configuring complex applications. Github, 2019. URL https://github.com/facebookresearch/hydra.

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*, pages 2223–2232, 2017.

Luisa Zintgraf, Kyriacos Shiarli, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *International Conference on Machine Learning*, pages 7693–7702. PMLR, 2019.

A Selecting ϵ for MNIST

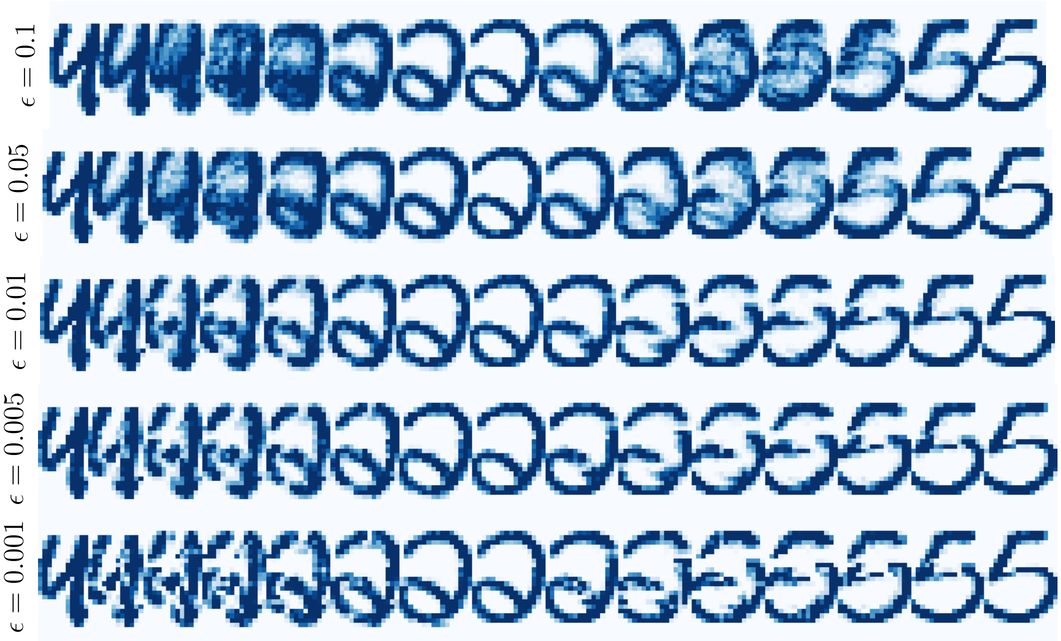

Figure 8: We selected $\epsilon = 10^{-2}$ for our MNIST coupling experiments as it results in transport maps that are not too blurry or sharp.

B Other models for continuous OT

While developing the hyper-network or Meta ICNN in sect. 3.2 for predicting couplings between continuous measures, we considered alternative modeling formulations briefly documented in this section. We finalized only the hyper-network model because it is conceptually the most similar to predicting the optimal dual variables in the continuous setting and results in rapid predictions.

B.1 Optimization-based meta-learning (MAML-inspired)

The model-agnostic meta-learning setup proposed in MAML (Finn et al., 2017) could also be applied in the Meta OT setting to learn an adaptable initial parameterization. In the continuous setting, one initial version would take a parameterized dual potential model $\psi_\varphi(x)$ and seek to learn an initial parameterization φ_0 so that optimizing a loss such as the W2GN loss $\mathcal{L}$ from eq. (12) results in a minimal $\mathcal{L}(\varphi_K)$ after adapting the model for K steps. Formally, this would optimize:

$$\arg\min_{\varphi_0} \mathcal{L}(\varphi_K) \quad \text{where} \quad \varphi_{t+1} = \varphi_t - \nabla_\varphi \mathcal{L}(\varphi_t) \tag{19}$$

Tancik et al. (2021) explores similar learned initializations for coordinate-based neural implicit representations for 2D images, CT scan reconstruction, and 3d shape and scene recovery from 2D observations.

Challenges for Meta OT. The transport maps given by $T = \nabla\psi$ can significantly vary depending on the input measures α, β. We found it difficult to learn an initialization that can be rapidly adapted, and optimizing eq. (19) is more computationally expensive than eq. (18) as it requires unrolling through many evaluations of the transport loss $\mathcal{L}$. And, we found that *only* learning to predict the optimal parameters with eq. (18), conditional on the input measures, and then fine-tuning with W2GN to be stable.

Advantages for Meta OT. Exploring MAML-inspired methods could further incorporate the knowledge that the model's prediction is going to be fine-tuned into the learning process. One promising direction we did not try could be to integrate some of the ideas from LEO (Rusu et al., 2018) and CAVIA (Zintgraf et al., 2019), which propose to learn a latent space for the parameters where the initialization is also conditional on the input.

B.2 Neural process and conditional Monge maps

The (conditional) neural process models considered in Garnelo et al. (2018b;a) can also be adapted for the Meta OT setting, and is similar to the model proposed in Bunne et al. (2022a). In the continuous setting, this would result in a dual potential that is also conditioned on a representation of the input measures, e.g. $\psi_\varphi(x; z)$ where $z := f^{\rm emb}_\varphi(\alpha, \beta)$ is a learned embedding of the input measures that is learned with the parameters of ψ. This could be formulated as

$$\operatorname*{arg\,min}_{\varphi} \mathop{\mathbb{E}}_{(\alpha,\beta)\sim\mathcal{D}} \mathcal{L}(\varphi, f^{\rm emb}_\varphi(\alpha, \beta)), \tag{20}$$

where $\mathcal{L}$ modifies the model used in the loss eq. (12) to also be conditioned on the context extracted from the measures.

Challenges for Meta OT. This raises the issue on best-formulating the model to be conditional on the context. One way could be to append z to the input point x in the domain. Bunne et al. (2022a) proposes to use the Partially Input-Convex Neural Network (PICNN) from (Amos et al., 2017) to make the model convex with respect to x and not z.

Advantages for Meta OT. A large advantage is that the representation z of the measures α, β would be significantly lower-dimensional than the parameters φ that our Meta OT models are predicting.

C Additional experimental and implementation details

We have attached the Jax source code necessary to run and reproduce all of the experiments in our paper and will open-source all of it. Here is a basic overview of the files:

```
├─meta_ot   Meta OT Python library code
│ ├─conjugate.py   Exact conjugate solver for the continuous setting
│ ├─data.py
│ ├─models.py
│ └─utils.py
├─config   Hydra configuration for the experiments (containing hyper-parameters)
├─train_discrete.py   Train Meta OT models for discrete OT
├─train_color_single.py   Train a single ICNN with W2GN between 2 images (for debugging)
├─train_color_meta.py   Train a Meta ICNN with W2GN
├─plot_mnist.py   Visualize the MNIST couplings
├─plot_world_pair.py   Visualize the spherical couplings
├─eval_color.py   Evaluate the Meta ICNN in the continuous setting
└─eval_discrete.py   Evaluate the Meta ICNN for the discrete tasks
```

Connecting to the data is one difficulty in running the experiments. The easiest experiment to re-run is the MNIST one, which will automatically download the dataset:

```
./train_discrete.py # Train the model, outputting to <exp_dir>
./eval_discrete.py <exp_dir> # Evaluate the learned models
./plot_mnist.py <exp_dir> # Produce further visualizations
```

C.1 Hyper-parameters

We briefly summarize the hyper-parameters we used for training, which we did not extensively tune. In the discrete setting, we use the same hyper-parameters for the MNIST and spherical settings.

Table 3: Discrete OT hyper-parameters.

Name	Value
Batch size	128
Number of training iterations	50000
MLP Hidden Sizes	[1024, 1024, 1024]
Adam learning rate	1e-3

Table 4: Continuous OT hyper-parameters.

Name	Value
Meta batch size (for α, β)	8
Inner batch size (to estimate $\mathcal{L}$)	1024
Cycle loss weight (γ)	3.
Adam learning rate	1e-3
ℓ_2 weight penalty	1e-6
Max grad norm (for clipping)	1.
Number of training iterations	200000
Meta ICNN Encoder	ResNet18
Encoder output size (both measures)	256×2
Meta ICNN Decoder Hidden Sizes	[512]

C.2 Sinkhorn convergence times, varying thresholds

In the main paper, table 1 reports the runtime of Sinkhorn to reach a convergence threshold of the marginal error being below a tolerance of 10^{23}. Tables 5 and 6 report the results from sweeping over other thresholds and show that Meta OT's initialization is consistently able to help.

Table 5: Sinkhorn runtime to reach a thresholded marginal error on MNIST.

Initialization	Threshold=10^{-2}	Threshold=10^{-3}	Threshold=10^{-4}	Threshold=10^{-5}
Zeros	$4.5 \cdot 10^{-3} \pm 1.5 \cdot 10^{-3}$	$7.7 \cdot 10^{-3} \pm 1.2 \cdot 10^{-3}$	$1.1 \cdot 10^{-2} \pm 1.8 \cdot 10^{-3}$	$1.5 \cdot 10^{-2} \pm 2.3 \cdot 10^{-3}$
Gaussian	$4.1 \cdot 10^{-3} \pm 1.2 \cdot 10^{-3}$	$7.7 \cdot 10^{-3} \pm 1.4 \cdot 10^{-3}$	$1.1 \cdot 10^{-2} \pm 1.7 \cdot 10^{-3}$	$1.4 \cdot 10^{-2} \pm 2.4 \cdot 10^{-3}$
Meta OT	$2.3 \cdot 10^{-3} \pm 9.2 \cdot 10^{-6}$	$3.9 \cdot 10^{-3} \pm 1.6 \cdot 10^{-3}$	$6.7 \cdot 10^{-3} \pm 1.4 \cdot 10^{-3}$	$1.0 \cdot 10^{-2} \pm 2.4 \cdot 10^{-3}$

Table 6: Sinkhorn runtime to reach a thresholded marginal error on the spherical transport problem.

Initialization	Threshold=10^{-2}	Threshold=10^{-3}	Threshold=10^{-4}	Threshold=10^{-5}
Zeros	$8.8 \cdot 10^{-1} \pm 1.3 \cdot 10^{-1}$	$1.4 \pm 1.9 \cdot 10^{-1}$	$2.1 \pm 3.6 \cdot 10^{-1}$	$2.8 \pm 5.6 \cdot 10^{-1}$
Gaussian	$5.6 \cdot 10^{-1} \pm 9.9 \cdot 10^{-2}$	$1.1 \pm 2.0 \cdot 10^{-1}$	$1.7 \pm 3.5 \cdot 10^{-1}$	$2.4 \pm 5.4 \cdot 10^{-1}$
Meta OT	$7.8 \cdot 10^{-2} \pm 3.4 \cdot 10^{-2}$	$0.44 \pm 1.5 \cdot 10^{-1}$	$0.97 \pm 3.2 \cdot 10^{-1}$	$1.7 \pm 6.8 \cdot 10^{-1}$

C.3 Experimental runtimes and convergence

App. C.3 shows the convergence during training of Meta OT models in the discrete and continuous settings over 10 trials on our single Quadro GP100 GPU. The MNIST models are consistently trained to optimality within 2 minutes (!) while the continuous model takes a few hours to train.

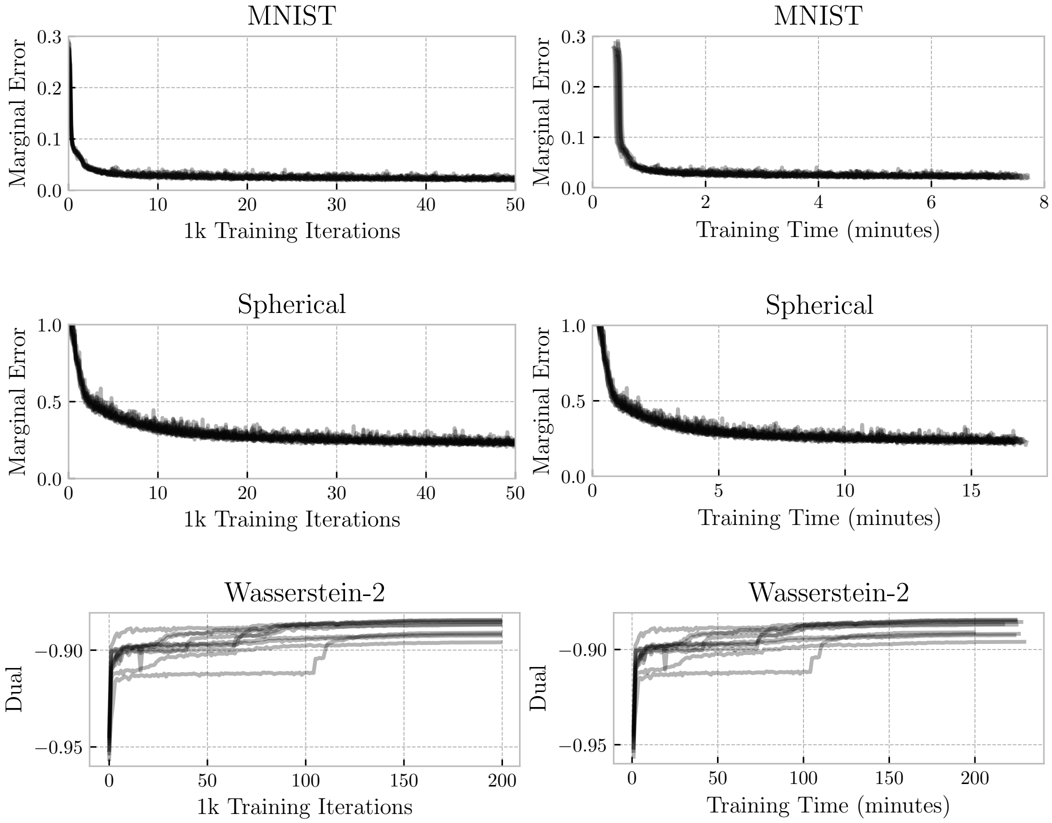

Figure 9: Convergence of Meta OT models during training, reported over iterations and wall-clock time. We run each experiment for 10 trials with different seeds and report each trial as a line.

D OUT-OF-DISTRIBUTION GENERALIZATION

App. D tests the ability of Meta OT to predict potentials for out-of-distribution input data. We consider the pairwise training and evaluation on the following datasets: 1) MNIST; 2) USPS (Hull, 1994) (upscaled to have the same size as the MNIST); 3) Google Doodles dataset* with classes Crab, Cat and Faces; 4) sparsified random uniform data in [0,1] where sparsity (zeroing values below 0.95) is used to mimic the sparse signal in black-and-white images. For each pair, eg, MNIST-USPS, we train on one dataset and use the other to predict the potentials. The comparison is done using the same metric as before, i.e., the deviation from the marginal constraints defined in eq. (7).

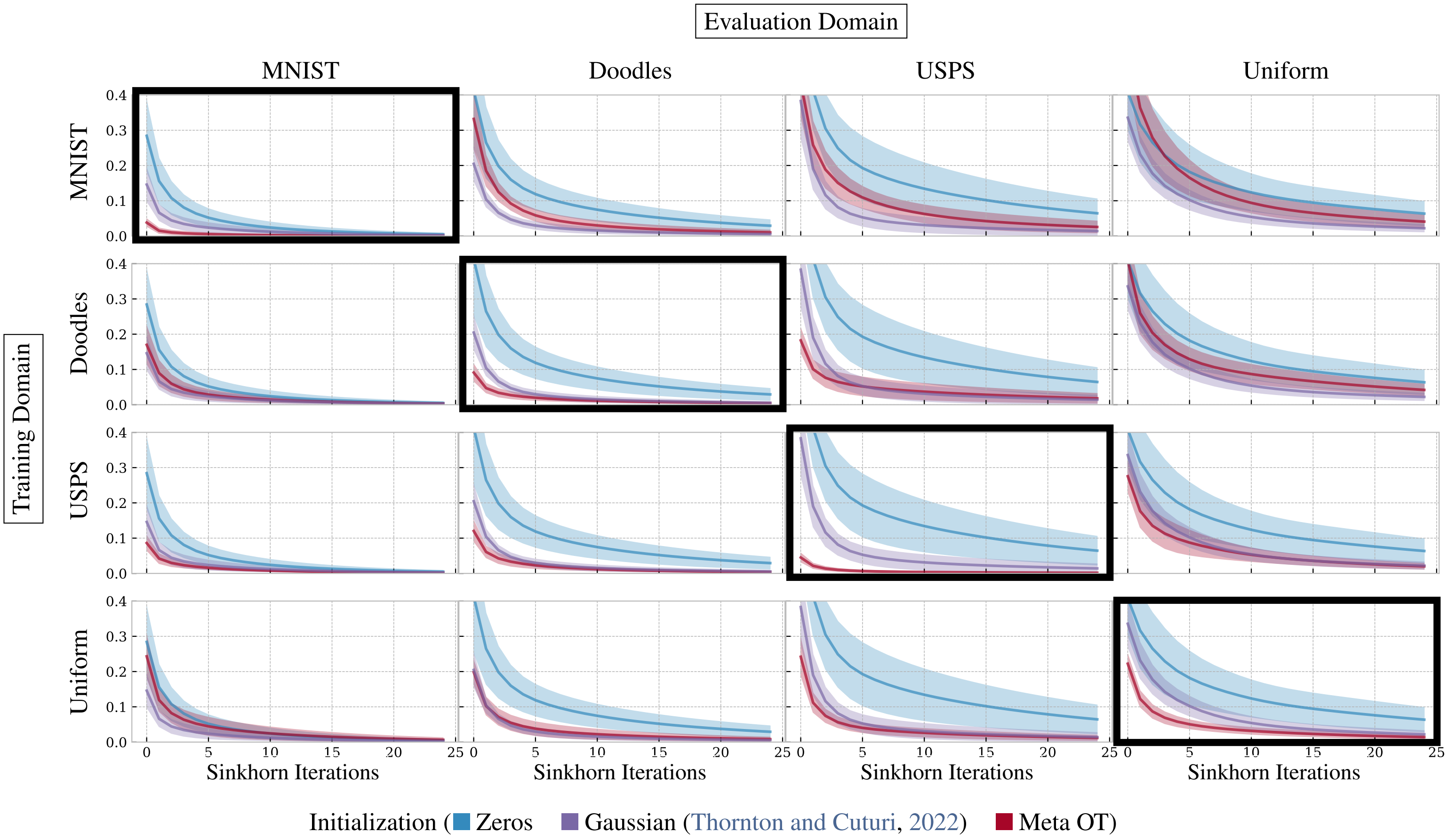

Figure 10: Cross-domain experiments.

*https://quickdraw.withgoogle.com/data

E Additional color transfer results

We next show additional color transfer results from the experiments in sect. 4.3 on the following public domain images from WikiArt:

- Distant View of the Pyramids by Winston Churchill (1921)
- Charing Cross Bridge, Overcast Weather by Claude Monet (1900)
- Houses of Parliament by Claude Monet (1904)
- October Sundown, Newport by Childe Hassam (1901)
- Landscape with House at Ceret by Juan Gris (1913)
- Irises in Monet's Garden by Claude Monet (1900)
- Crystal Gradation by Paul Klee (1921)
- Senecio by Paul Klee (1922)
- Váza s květinami by Josef Capek (1914)
- Sower with Setting Sun by Vincent van Gogh (1888)
- Three Trees in Grey Weather by Claude Monet (1891)
- Vase with Daisies and Anemones by Vincent van Gogh (1887)

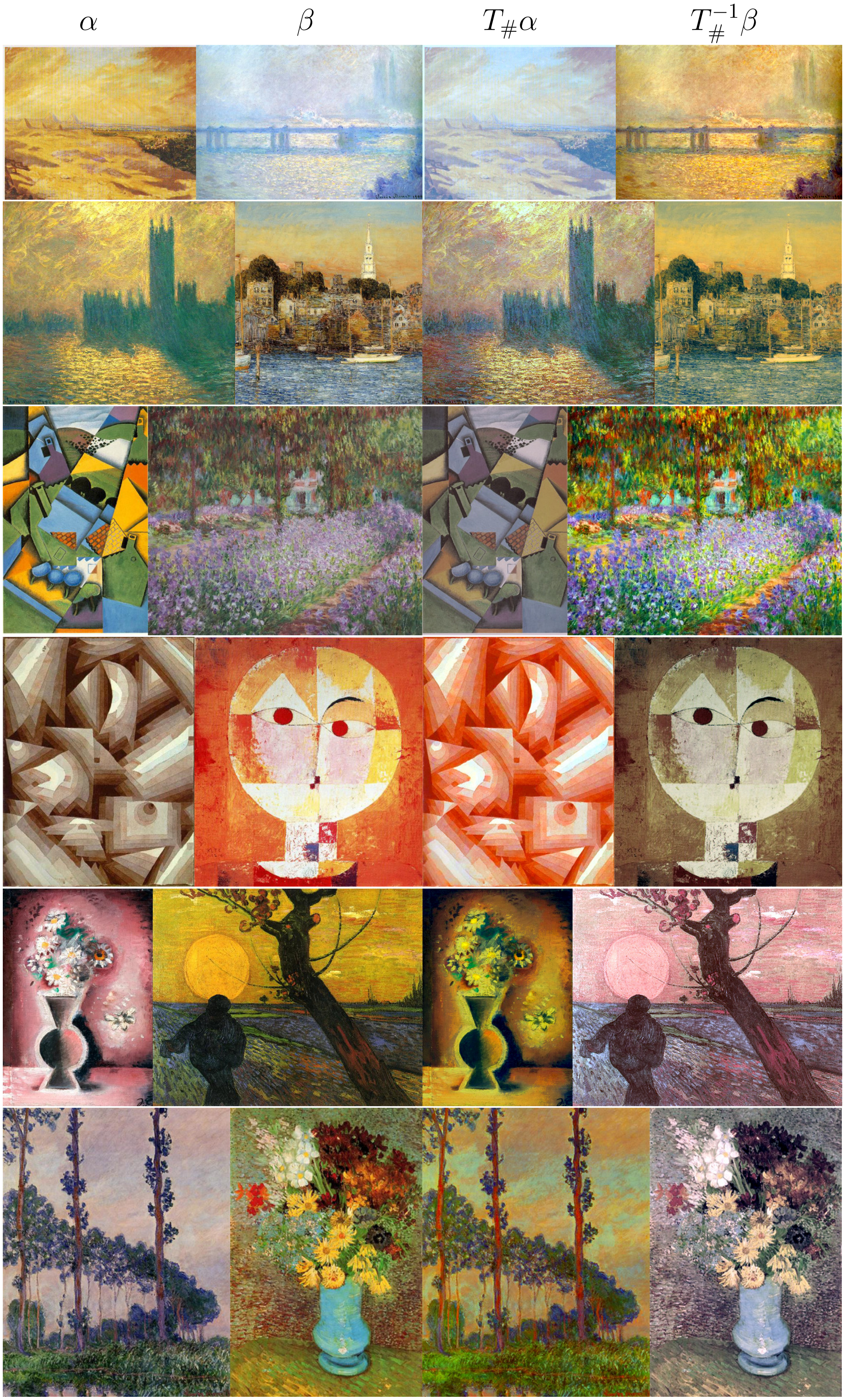

Figure 11: Meta ICNN (initial prediction). The sources are given in the beginning of app. E.

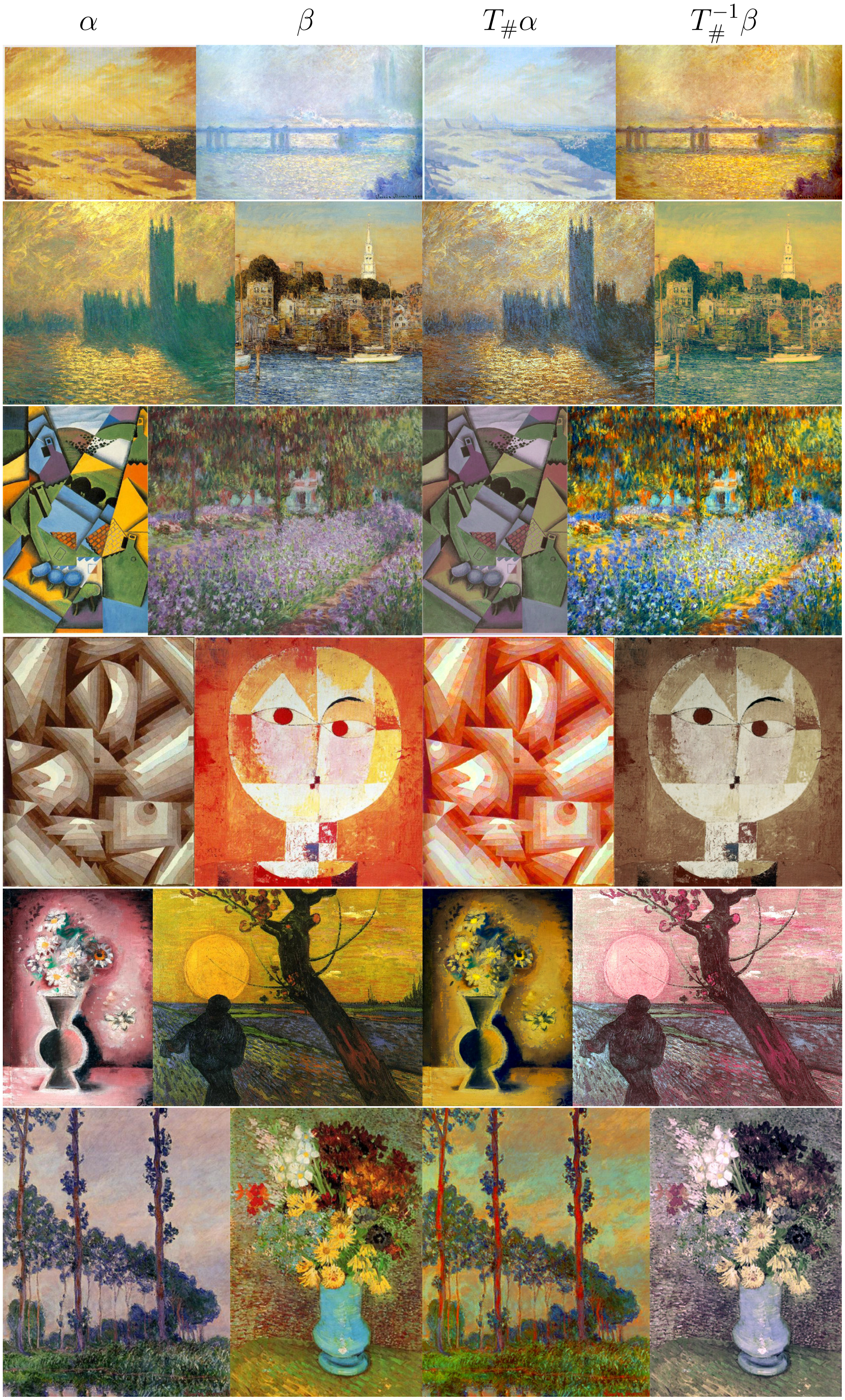

Figure 12: Meta ICNN + W2GN fine-tuning. The sources are given in the beginning of app. E.

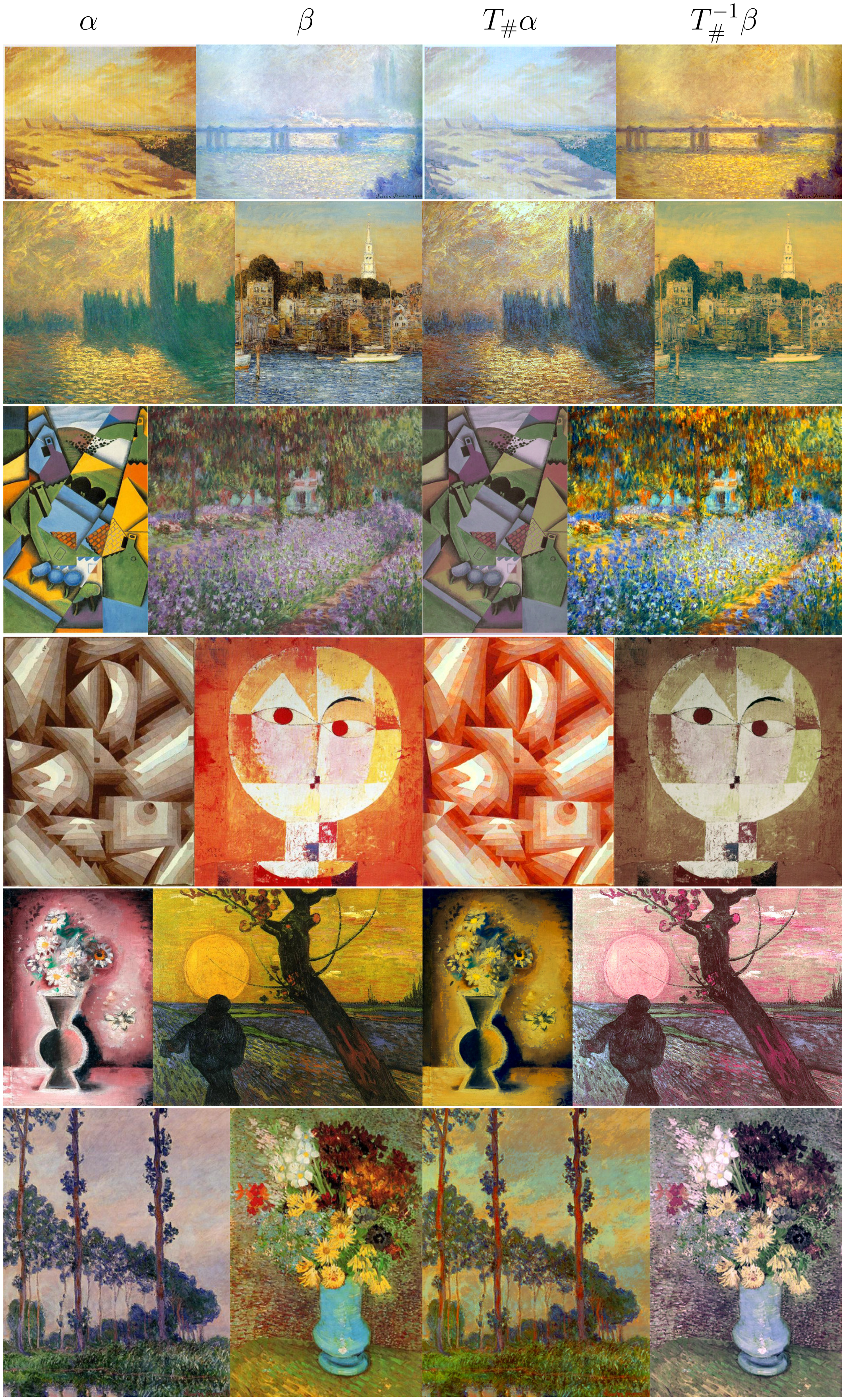

Figure 13: W2GN (final). The sources are given in the beginning of app. E.