SyMOT-Flow: Learning optimal transport flow for two arbitrary distributions with maximum mean discrepancy

Zhe Xiong*
School of Mathematical Sciences
Shanghai Jiao Tong University
Shanghai 200240, China
aristotle-x@sjtu.edu.cn

Qiaoqiao Ding
Institute of Natural Sciences
Shanghai Jiao Tong University
Shanghai 200240, China
dingqiaoqiao@sjtu.edu.cn

Xiaoqun Zhang
Institute of Natural Sciences
Shanghai Jiao Tong University
Shanghai 200240, China
xqzhang@sjtu.edu.cn

Abstract

Finding a transformation between two unknown probability distributions from samples is crucial for modeling complex data distributions and perform tasks such as density estimation, sample generation, and statistical inference. One powerful framework for such transformations is normalizing flow, which transforms an unknown distribution into a standard normal distribution using an invertible network. In this paper, we introduce a novel model called SyMOT-Flow that trains an invertible transformation by minimizing the symmetric maximum mean discrepancy between samples from two unknown distributions, and we incorporate an optimal transport cost as regularization to obtain a short-distance and interpretable transformation. The resulted transformation leads to more stable and accurate sample generation. We establish several theoretical results for the proposed model and demonstrate its effectiveness with low-dimensional illustrative examples as well as high-dimensional generative samples obtained through the forward and reverse flows.

Finding a transformation between two unknown probability distributions from samples has many applications in machine learning and statistics, for example density estimation [1] and sample generation [2, 3], for both we can use the transformation to generate new samples from the target distribution. Furthermore, finding a transformation between two unknown probability distributions can also be useful in domains such as computer vision, speech recognition, and natural language processing, where we often encounter complex data distributions. For example, in computer vision, we can use the transformation to model the distribution of images and generate new images with desired characteristics [4, 5].

There are several common techniques for finding the transformation between two probability distributions. Normalizing flow (NF) is a popular and powerful modeling technique which has attracted significant attention in statistics and machine learning fields [6]. Normalizing flow involves defining a sequence of invertible transformations between probability distributions, where each transformation

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Accepted to the Optimal Transport and Machine Learning Workshop at NeurIPS 2023.

is designed to be easy to compute and invert. By applying a sequence of such transformations to a simple distribution, such as a Gaussian distribution, we can generate more complex distributions that can be used to model complex datasets. On this purpose, the structure of NF needs to be elaborately designed such that the transformation is invertible and the Jacobian determinant is tractable [7, 8, 9]. As a widely used generative model, it has a good performance for both sampling and density evaluation tasks [10, 11, 7].

On the other hand, optimal transport (OT) [12, 13] is a classical mathematical framework involving finding the optimal mapping between two probability distributions that minimizes a cost function, such as the Wasserstein distance [14, 15, 16]. Optimal transport has been combined with different generative models to improve the quality and stability of the generated samples [17, 18, 19, 20, 21].

Combined with deep learning, a few works have been proposed to learn the transformation between two sets of samples. Coeurdoux et al. [17] proposed SWOT-Flow to use the sliced-Wasserstein distance as the distance between the transformed distribution to the objective one. Also, in their work, they used normalizing flow to approximate the transformation and add optimal transport and Jacobian regularization to improve the performance. Besides, in [22] three invertible flow models are combined together by maximizing the likelihood of both distributions respectively. The choice of distance measure between two probability distributions is essential to the performance and characteristic of generative models. For example Kullback-Leibler (KL) divergence and the Wasserstein distance are adopted in Generative Adversarial Networks (GANs). In the case of continuous distributions, the maximum log-likelihood in NF is equivalent to minimizing the KL divergence between the transformed distribution and the normal distribution. However, recent research has explored the use of alternative distance metrics, such as the Kernel Stein Discrepancy (KSD) [23, 24], for posterior approximation in generative models. These alternative metrics offer new opportunities for improving the accuracy and efficiency of generative models in various applications [25, 26]. Maximum mean discrepancy (MMD) [27] is another important metrics to find a continuous function to give the difference in mean values between samples. In [28, 29], the discriminator in GAN is replaced by MMD between the generated and data points.

Motivated by invertible transformation constructed in normalizing flow approaches, in this paper, we propose a method to learn an invertible transformation between two unknown distributions based on given samples, namely **SyM**metrical **M**MD **OT**-**Flow** (SyMOT-Flow). In our model, the two-direction maximum mean discrepancy (MMD) [27] is used to measure the discrepancy between the transformed samples to the original ones. Besides, we consider the consistency to OT and add the OT cost in Monge's problem [12] as a regularization. Inspired by the theoretical work derived in [30, 31], we demonstrate the existence and feasibility of our proposed model. We also analyze the properties of solutions with respect to the regularization parameter. The proposed model takes the advantages of kernel in MMD for capturing intrinsic structure of samples and the regularity and stability of parameterized optimal transport. Finally the transformation is constructed through a sequence of invertible network structure which enable continuity and invertibility between two distributions and high dim datasets. In feature space, the learn transformation allows an optimal correspondence of samples, which can be used for further applications such as generative modeling, feature matching and domain adaptation. Extensive experiments on both low-dimension illustrative examples and data-sets demonstrate the performance of our model. Also, ablation studies on the effect of the OT regularization and symmetrical designs of our models are provided to show the characteristic of learned transformation.

This paper is organized as follows. Section 1 gives the notation and preliminary. Section 2 describes the proposed method and gives the theoretical results. Section 3 is devoted to the experimental evaluation and comparison to other methods. Section 4 concludes the paper.

1 Notations

In the following, we introduce some notations and background for the proposed method. For simplicity, we consider p and q as two unknown distributions defined in the space $\mathbb{R}^d$. Suppose $\mathbf{x}$, $\mathbf{z}$ are two random variables with distribution p, q. Correspondingly, $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{z}_j\}_{j=1}^{N'}$ are samples from p and q respectively.

Maximum Mean Discrepancy (MMD) Suppose $\mathcal{F} := \left\{ f : f : \mathbb{R}^d \to \mathbb{R} \right\}$ is a class of functions. Then the MMD between p and q is defined as:

$$\mathrm{MMD}(\mathcal{F}, p, q) = \sup_{f \in \mathcal{F}} \mathbb{E}_p \left[f(\mathbf{x}) \right] - \mathbb{E}_q \left[f(\mathbf{z}) \right].$$

Specially, the function class $\mathcal{F}$ is chosen to be an RKHS space $\mathcal{H}$ and the corresponding squared MMD is reformulated as

$$\mathrm{MMD}(p, q)^2 = \|\mu_p - \mu_q\|_{\mathcal{H}}^2,$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm of space $\mathcal{H}$ and $\mu_p \in \mathcal{H}$ is the mean embedding of p given by

$$\mu_p = \int_{\mathcal{X}} k(\mathbf{x}, \cdot) p(d\mathbf{x}) \in \mathcal{H},$$

and $k(\cdot, \cdot)$ is a kernel function defined on the space $\mathcal{H} \times \mathcal{H}$. Let $\phi : \mathbb{R}^d \to \mathbb{R}^{d_\phi}$ be the feature map associated with $k(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^\top \phi(\mathbf{z})$, then the squared MMD can be simplified as

$$\mathrm{MMD}(p, q)^2 = \|\mathbb{E}_{\mathbf{x} \sim p} \left[\phi(\mathbf{x}) \right] - \mathbb{E}_{\mathbf{z} \sim q} \left[\phi(\mathbf{z}) \right] \|_2^2. \tag{1}$$

Moreover, the MMD can also be represented by the kernel function $k(\cdot, \cdot)$:

$$\mathrm{MMD}(p, q)^2 = \mathbb{E}_{\mathbf{x} \sim p, \mathbf{x}' \sim p} \left[k(\mathbf{x}, \mathbf{x}') \right] + \mathbb{E}_{\mathbf{z} \sim q, \mathbf{z}' \sim q} \left[k(\mathbf{z}, \mathbf{z}') \right] - 2\mathbb{E}_{\mathbf{x} \sim p, \mathbf{z} \sim q} \left[k(\mathbf{x}, \mathbf{z}) \right]. \tag{2}$$

Empirically, for the samples $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{z}_j\}_{j=1}^{N'}$, we have the discrete estimator of MMD:

$$\mathrm{MMD}_b(p, q)^2 = \frac{1}{NN'} \sum_{n=1}^{N} \sum_{n'=1}^{N'} \left[k(\mathbf{x}_n, \mathbf{x}_{n'}) - 2k(\mathbf{x}_n, \mathbf{z}_{n'}) + k(\mathbf{z}_n, \mathbf{z}_{n'}) \right]. \tag{3}$$

Optimal Transport (OT) Suppose $c(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is a nonnegative cost function, then the optimal transport problem by Monge [12] is given by

$$\min_T \int_{\mathcal{X}} c(\mathbf{x}, T(\mathbf{x})) dp(\mathbf{x}) \quad s.t.\ T_\sharp p = q, \tag{4}$$

where $T : \mathbb{R}^d \to \mathbb{R}^d$ is a measurable mapping and $T_\sharp$ is the push-forward operator such that

$$[T_\sharp p = q] \Longleftrightarrow \left[\int_{\mathbb{R}^d} h(\mathbf{z}) dq(\mathbf{z}) = \int_{\mathbb{R}^d} h(T(\mathbf{x})) dp(\mathbf{x}), \forall h \in \mathcal{C}_0(\mathbb{R}^d) \right].$$

where $\mathcal{C}_0(\mathbb{R}^d)$ means the space of continuous functions vanishing at infinity on $\mathbb{R}^d$. We can obtain a relaxed OT form of problem (4) on replacing the equality $T_\sharp p = q$ by a distribution distance $d(\cdot, \cdot)$ as

$$\min_T \int_{\mathcal{X}} c(\mathbf{x}, T(\mathbf{x})) dp(\mathbf{x}) + \lambda d(T_\sharp p, q), \tag{5}$$

where $\lambda > 0$ is the weight of the distance penalty.

Invertible Neural Networks (INNs) Invertible Neural Networks (INNs) are neural networks architectures with invertibility by design, which are often composed of invertible modules such as affine coupling layers [8] or neural ODE [32]. With these specially designed structures, it tends to be tractable to compute the inverse transformation and Jacobian determinant, which is widely used in the NF tasks. In the coupling layer, the input $\mathbf{x}$ is split along the channels into two part $(\mathbf{x}_1, \mathbf{x}_2)$ and then is transformed as follows,

$$\begin{aligned} \mathbf{z}_1 &= \mathbf{x}_1, \\ \mathbf{z}_2 &= \mathbf{x}_2 \odot \exp\left(\gamma * \tanh(\mathbf{s}_{\theta_1}(\mathbf{x}_1))\right) + \mathbf{t}_{\theta_2}(\mathbf{x}_1), \end{aligned}$$

where γ is the affine clamp parameter. Here $\mathbf{s}_{\theta_1}(\cdot)$ and $\mathbf{t}_{\theta_{2}(\cdot)}$ are two subnets to be trained, whose structures can be different [8] or the same [9]. In the last step, the output $(\mathbf{z}_1, \mathbf{z}_2)$ are primarily concatenated and then disordered along the channels by an 1×1 invertible convolutional transform.

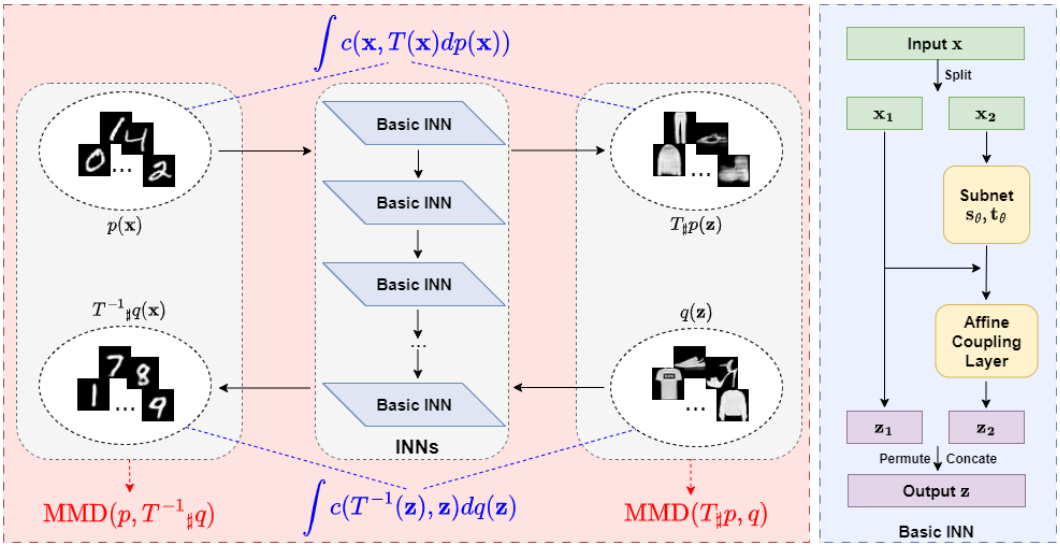

Figure 1: Overview of SyMOT-Flow Model.

2 Our method

The diagram of our model is presented in Fig 1. As mentioned in Section 1, we choose $d(\cdot, \cdot)$ to be the squared MMD in the relaxed OT (5). Moreover, to improve the stability of the transformation T, we make use of the invertibility of T and design a symmetrical distance as follows:

$$d_{\mathrm{MMD}}(T, p, q) = \mathrm{MMD}(T_\sharp p, q)^2 + \mathrm{MMD}(p, T^{-1}{}_\sharp q)^2.$$

Correspondingly, we also add a symmetrical cost to the objective function in OT and finally the loss function with parameter λ is defined as

$$L_T = \left(\int_{\mathbb{R}^d} c(\mathbf{x}, T(\mathbf{x}))dp(\mathbf{x}) + \int_{\mathbb{R}^d} c(T^{-1}(\mathbf{z}), \mathbf{z})dq(\mathbf{z}) \right) + \lambda d_{\mathrm{MMD}}(T, p, q). \tag{6}$$

In practice, suppose T_θ is an invertible network with parameters θ. Given two sets of samples $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{z}_j\}_{j=1}^{N'}$. The empirical training loss function is defined as follows:

$$\begin{aligned} L_\theta = & \frac{1}{NN'} \sum_{n=1}^{N} \sum_{n'=1}^{N'} \left[k(T_\theta(\mathbf{x}_n), T_\theta(\mathbf{x}_{n'})) + k(T_\theta^{-1}(\mathbf{z}_n), T_\theta^{-1}(\mathbf{z}_{n'}) \right] \\ & - \frac{2}{NN'} \sum_{n=1}^{N} \sum_{n'=1}^{N'} \left[k(T_\theta(\mathbf{x}_n), \mathbf{z}_{n'}) + k(\mathbf{x}_n, T_\theta^{-1}(\mathbf{z}_{n'})) \right] \\ & + \frac{\beta}{N} \sum_{i=1}^{N} c(\mathbf{x}_i, T_\theta(\mathbf{x}_i)) + \frac{\beta}{N'} \sum_{j=1}^{N'} c(T_\theta^{-1}(\mathbf{z}_j), \mathbf{z}_j), \end{aligned} \tag{7}$$

where $\beta = \frac{1}{\lambda} > 0$ is the weight parameter and T_θ^{-1} is the inversion of T_θ. Note that the empirical loss (7) is slightly different from (6) as we put the weight on the MMD term for actual implementation. As opposed to merely minimizing the OT cost, it is crucial to prioritizing the attainment of a close-to-zero MMD to establish the validity of the constraint $T_\sharp p = q$. Moreover, in the calculation of empirical MMD, we omit two items which are irrelevant to the parameters of the invertible network. Before introducing theoretical results of our models, we firstly make some assumptions as follows:

Assumption 2.1. *The numbers of samples N and N' from distributions p and q are both large enough to make sure that $P\left\{|\mathrm{MMD}_b(p, q)^2 - \mathrm{MMD}(p, q)^2| > \delta\right\} < \epsilon$ for small numbers δ and ϵ, which implies that the empirical estimator MMD_b is closed to the true MMD between p and q such that the transformation T_θ obtained from the samples is an accurate approximation to the theoretical solution T.*

Assumption 2.2. *The optimal transport problem* (4) *is solved in the space of invertible and continuous functions, which means that the optimal plan T between the distribution p and q exists and is an invertible and continuous function.*

Assumption 2.2 ensures that the optimal transport T is invertible, therefore in our theorem we consider the symmetrical OT problem and squared MMD. Recall the problem (4) and the symmetrical version is defined as

$$\min_{T} \left\{ \int_{\mathbb{R}^d} c(\mathbf{x}, T(\mathbf{x}))dp(\mathbf{x}) + \int_{\mathbb{R}^d} c(T^{-1}(\mathbf{z}), \mathbf{z})dq(\mathbf{z}) : T_\sharp p = q. \right\} := \mathrm{OT}(p, q), \tag{8}$$

Correspondingly, the relaxation Monge's problem is defined as

$$\left\{ \min_{T} \int_{\mathbb{R}^d} c(\mathbf{x}, T(\mathbf{x}))dp(\mathbf{x}) + \int_{\mathbb{R}^d} c(T^{-1}(\mathbf{z}), \mathbf{z})dq(\mathbf{z}) + \lambda d_{\mathrm{MMD}}(T, p, q) \right\} := \mathrm{OT}_\lambda(p, q). \tag{9}$$

Assumption 2.3 (Existence of solution to relaxed OT)**.** *For any $\lambda > 0$, the relaxed Monge's problem* (9) *always has an optimal plan $T^\star_\lambda$, which is invertible and continuous.*

As we mentioned above, now our goal is to deal with the problem (9) and the following theorem reveals the relationship between the optimal solutions of problem (8) and (9):

Theorem 2.1. *Suppose p and q are two probability measures in $\mathbb{R}^d$, where d is the dimension of variables $\mathbf{x}$ and $\mathbf{z}$. If the kernel function k is continuous and integrally strictly positive definitive (integrally s.p.d) such that the corresponding RKHS $\mathcal{H} \subset \mathcal{C}_0$, then for any positive and increasing sequence $\{\lambda\}$, it holds that,*

$$\lim_{\lambda \to +\infty} \mathrm{OT}_\lambda(p, q) = \mathrm{OT}(p, q). \tag{10}$$

Proof of Theorem 2.1. Suppose for each $\lambda > 0$, $T^\star_\lambda$ is an minimizer of problem $\mathrm{OT}_\lambda(p, q)$ and $T^\star$ is the minimizer of the original OT problem $\mathrm{OT}(p, q)$ respectively. By the definition of minimizer, for $T^\star_\lambda$ we have that

$$\begin{aligned} \mathrm{OT}(p, q) &= \int_{\mathbb{R}^d} c(\mathbf{x}, T^\star(\mathbf{x}))dp(\mathbf{x}) + \int_{\mathbb{R}^d} c(T^{\star -1}(\mathbf{z}), \mathbf{z})dq(\mathbf{z}) \\ &\geq \int_{\mathcal{X}} c(\mathbf{x}, T^\star_\lambda(\mathbf{x}))dp(\mathbf{x}) + \int_{\mathbb{R}^d} c(T^{\star -1}_\lambda(\mathbf{z}), \mathbf{z})dq(\mathbf{z}) + \lambda d_{\mathrm{MMD}}(T^\star_\lambda, p, q) = \mathrm{OT}_\lambda(p, q). \end{aligned} \tag{11}$$

Then consequently, it holds that

$$\limsup_{\lambda \to +\infty} \lambda d_{\mathrm{MMD}}(T^\star_\lambda, p, q) \leq \limsup_{\lambda \to +\infty} \mathrm{OT}_\lambda(p, q) \leq \mathrm{OT}(p, q) < +\infty. \tag{12}$$

On the other hand, from inequality (12) it is easy to get that

$$\lim_{\lambda \to +\infty} d_{\mathrm{MMD}}(T^\star_\lambda, p, q) = 0. \tag{13}$$

According to [33, Lemma 3], since the kernel function k is integrally s.p.d and $\mathcal{H} \subset \mathcal{C}_0$, the limit (13) indicates that $T^\star_{\lambda\,\sharp} p \to q$ and ${T^\star_{\lambda}}_\sharp^{-1} q \to p$ in weak sense. Hence we can get the result that

$$\liminf_{\lambda \to +\infty} \mathrm{OT}_\lambda(p, q) \geq \liminf_{\lambda \to +\infty} \int_{\mathcal{X}} c(\mathbf{x}, T^\star_\lambda(\mathbf{x}))dp(\mathbf{x}) + \int_{\mathbb{R}^d} c(T^{\star -1}_\lambda(\mathbf{z}), \mathbf{z})dq(\mathbf{z}) = \mathrm{OT}(p, q), \tag{14}$$

where the last equality comes from the weak convergence from $T^\star_{\lambda\,\sharp} p$ to q and from ${T^\star_{\lambda}}_\sharp^{-1} q$ to p. Then combining the results of (12) and (14) we conclude that

$$\lim_{\lambda \to +\infty} \mathrm{OT}_\lambda(p, q) = \mathrm{OT}(p, q).$$

□

Referring to [30, Theorem 1], the next theorem guarantees the existence of the solution to the MMD relaxation for the OT problem, which corroborates the feasibility of using MMD as the distribution distance.

Theorem 2.2. *For any $\epsilon > 0$, there exists a K and a series of invertible transformations $\{T_i\}_{i=1}^K$ such that $T = T_K \circ \cdots \circ T_1$ and*

$$\mathrm{MMD}(T_\sharp p, q) + \mathrm{MMD}(p, (T^{-1})_\sharp q) < \epsilon. \tag{15}$$

Proof of Theorem 2.2. Similar to the demonstration in [30, Lemma 1], with the conditions in the theorem we obtain that

$$\mathrm{MMD}(T_\sharp p, q) < \epsilon. \tag{16}$$

On the other hand, by the original definition of $\mathrm{MMD}(p, T_\sharp^{-1} q)$ and letting Hilbert space $\mathcal{H}$ contains the invertible and continuous transformations, we can obtain that

$$\begin{aligned}
\mathrm{MMD}(p, T_\sharp^{-1} q) &= \sup_{f \in \mathcal{H}} \mathbb{E}_p[f(\mathbf{x})] - \mathbb{E}_{T_\sharp^{-1} q}[f(\mathbf{y})] \\
&= \sup_{f \in \mathcal{H}} \mathbb{E}_{T_\sharp p}[f(T^{-1}(\mathbf{x}))] - \mathbb{E}_q[f(T^{-1}(\mathbf{y}))] \\
&\leq \sup_{f \in \mathcal{H}} \mathbb{E}_{T_\sharp p}[f(\mathbf{x})] - \mathbb{E}_q[f(\mathbf{y})] \\
&= \mathrm{MMD}(T_\sharp p, q) < \epsilon.
\end{aligned} \tag{17}$$

Therefore, with simple scaling we finally get that

$$\mathrm{MMD}(T_\sharp p, q) + \mathrm{MMD}(p, T_\sharp^{-1} q) < \epsilon. \tag{18}$$

□

With Theorem 2.2, it is feasible to get an optimal transformation between p and q through a series of invertible normalizing flow modules such as RealNVP[8], Glow[9], etc.

3 Experiments

Implementation details. For all the experiments, we use FrEIA [34] to build invertible flow structure. Moreover, in the calculation of MMD_b, we use multi-kernel MMD,i.e. a weighted combination of multiple kernels [35, Section 2.4]. For all the experiments, the flow contains 8 INN blocks with fully connected or convolutional subnets. The batch size is equal to 200 and the optimizer is chosen as Adam.

For each pair of two datasets, the weight parameter β in the empirical loss (7) is a hyperparameter which has been fine-tuned for the best performance. Besides, we also have some ablation experiments about the influence of weight β and the symmetrical design to the performance of the optimal solution learned by INN.

Illustrative examples. To validate the effectiveness of the proposed approach, we simulate four pairs of illustrative examples in $\mathbb{R}^2$. The proposed SyMOT-Flow method is compared with the model only with MMD (single MMD with $\beta = 0$ in (7)) and SWOT-Flow proposed in Coeurdoux et al. [17]. Fig.2 shows two sets of experiments. In all the experiments, we randomly pick 2000 sample points $\{\mathbf{x}_i\}$ and $\{\mathbf{z}_j\}$ from each distribution, represented by blue points in the two rows respectively. For the first one, the two sets of points are drawn from a series of Gaussian distributions with the centers along two lines with different covariance respectively. And for the second one, the data are sampled from two-moon and circles, with noise variance 0.05.

In each sub-figure of Fig.2, the blue points represent the original samples and the orange ones represent the generated data $\{T_\theta(\mathbf{x}_i)\}$ and $\{{T_\theta}^{-1}(\mathbf{z}_j)\}$ via the learned mappings T_θ and T_θ^{-1} (the correspondence are linked with green lines) by different methods. It can be seen that with the OT regularization, the method tends to learn a map which gives the shortest distance from one sets to the another, comparing to the single MMD and SWOT-Flow methods. Moreover, the symmetric loss provides a more stable sampling for both forward and inverse transformation.

Numerically, the corresponding OT cost and MMD distances are shown in Table 1 for each method and each pair of data. SyMOT-Flow obtains smaller OT cost and MMD distance for forward and backward of flow in comparison with single MMD and SWOT-Flow, as expected.

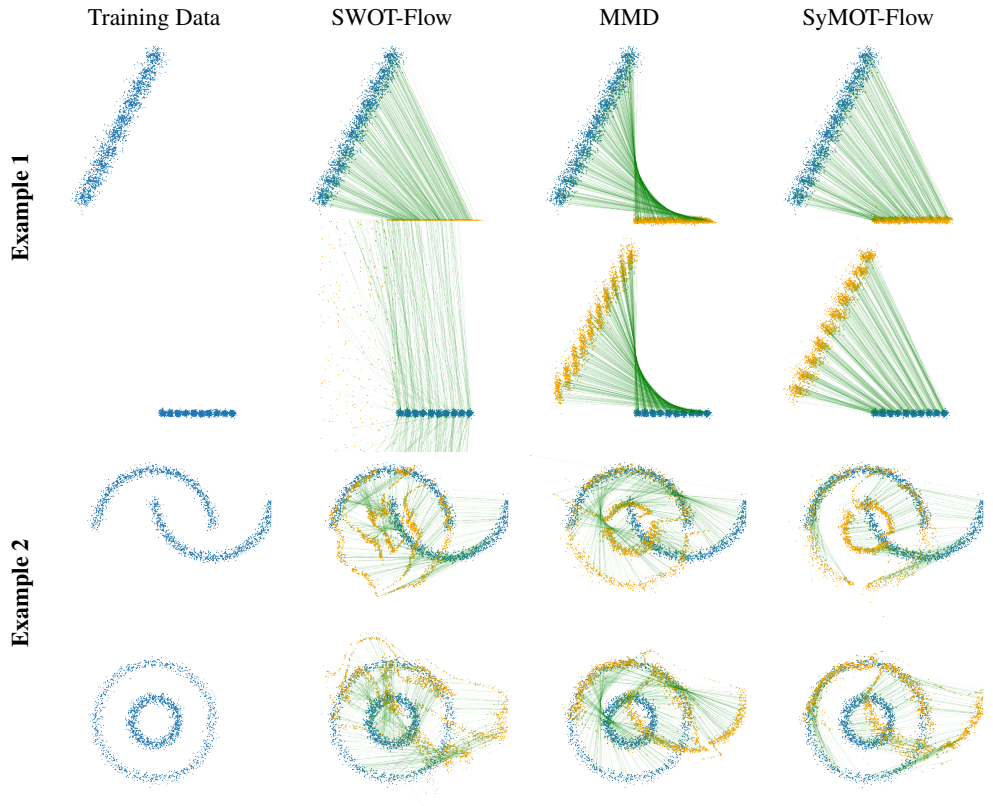

Figure 2: The input (blue) and generated points (orange) by learned map (green lines) between two sets of training data with different methods.

Table 1: The OT cost and MMD distance of forward/backward direction of flow.

	Method	SWOT-Flow	single MMD	SyMOT-Flow
OT Cost	Moons to Circles	0.684/0.723	0.924/0.889	**0.295/0.288**
	Gauss to Gauss	32.924/32.940	35.310/33.956	**32.459/32.475**
	8 Gauss to 8 Gauss	7.903/7.924	17.427/17.799	**7.539/7.483**
	Linear Gauss	**138.580**/5834.745	156.393/153.093	139.780/**134.318**
MMD distance	Moons to Circles	1.7e-2/4.3e-2	6.3e-3/5.7e-3	**2.9e-3/2.7e-3**
	Gauss to Gauss	6.6e-2/6.6e-2	**2.3e-2/1.8e-2**	6.6e-2/6.3e-2
	8 Gauss to 8 Gauss	1.4e-2/1.4e-2	7.9e-3/3.9e-3	**4.2e-3/2.5e-3**
	Linear Gauss	5.6e-3/3.5	3.3e-3/7.0e-3	**1.1e-3/1.3e-3**

MINST and Fashion-MINST. We evaluate SyMOT-Flow on two sets of high dimension data: MNIST and Fashion-MNIST. We pretrain an auto-encoder-decoder with MNIST and Fashion-MNIST dataset respectively. Then, SyMOT-Flow is applied to learn the transformation in the feature spaces. In the inference process, taking the transfer from MNIST to Fashion-MNIST as example, the working flow is to apply the encoder of MNIST, SyMOT-Flow and the decoder of Fashion-MNIST. Fig. 3 show the results of the generated samples between MNIST and Fashion-MNIST datasets with the learned transformation of SyMOT-Flow. It can be seen that SyMOT-Flow can generate high quality data in high dimension space.

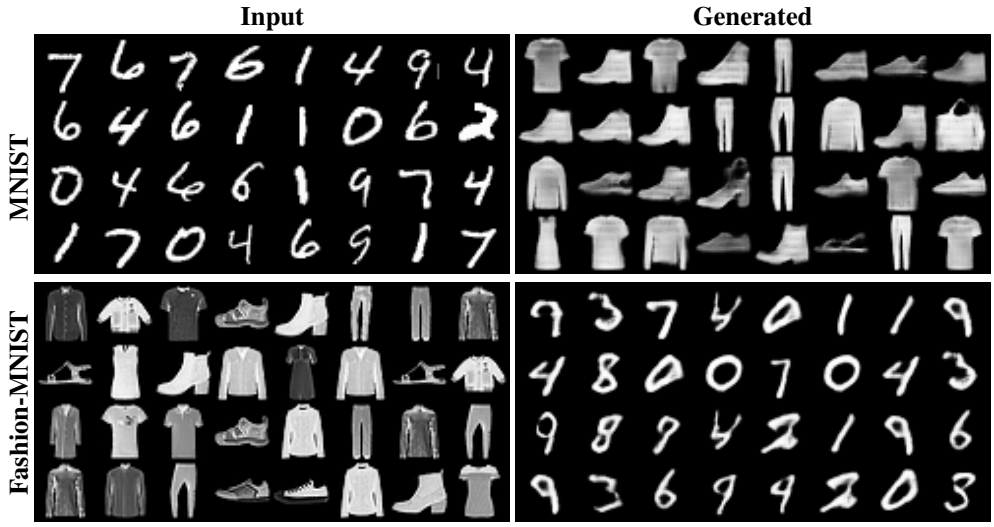

Figure 3: The generation results between MNIST and Fashion-MNIST datasets

Ablation Study. To assess the impact of the symmetric reverse flow component, experiments are conducted and the results are shown in Fig. 4 for the case with and without the reversed flow loss. It can be seen that in the absence of the reversed component, the generated samples exhibits a higher level dissimilarity to the intrinsic distribution, compared to those generated with the proposed symmetrical loss.

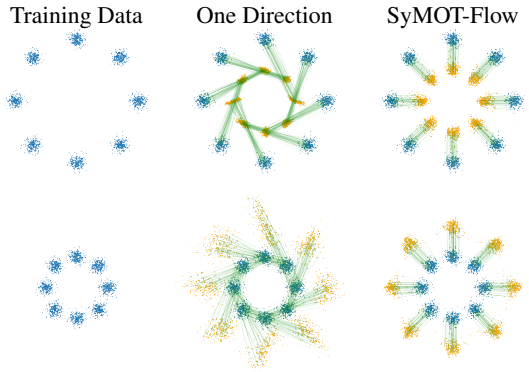

Figure 4: The input $\mathbf{x}$ (blue) and generated points $\mathbf{z}$ (orange) by learned map T_θ (green lines) between two sets of training data with and without the reversed flow loss.

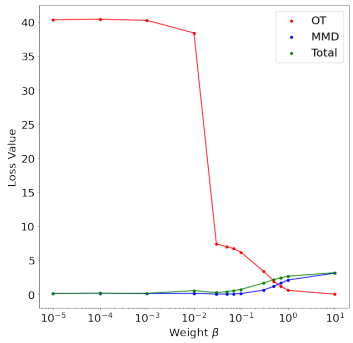

Figure 5: The value of MMD distance and OT cost with increasing weight parameter β.

The selection of the weight parameter β for the OT regularization plays a crucial role in achieving superior performance in learning the optimal transport. The change of values of MMD and OT referring to several different β is shown in Figure 5 and the corresponding OT costs and MMD distances are presented in Table 2. The red line corresponds to the values of the OT cost, the blue line represents the results of the MMD distance, and the green line indicates the total loss. The weight parameter β varies from 10^{-5} to 10 with 10 logarithmically spaced increments, more refined results are displayed. It is obvious as β increases, the OT cost decreases. Meanwhile, it is crucial to maintain a close-to-zero MMD throughout this process, as an excessively large β would result in the learned transformation being an identity map. More visualized results are provided in supplementary.

Table 2: The final values of OT cost and MMD distance for different weight β.

Weight β	1e-5	1e-4	1e-3	1e-2	1e-1	1e0	1e1
OT	40.321	40.390	40.228	38.501	6.134	0.577	**0.006**
MMD	0.124	0.142	0.097	0.256	**0.096**	2.071	3.099
Weight β	3e-2	5e-2	7e-2	1e-1	3e-1	5e-1	7e-1
OT	7.371	7.000	6.694	6.163	3.392	1.940	**1.142**
MMD	**0.016**	0.031	0.050	0.095	0.603	1.158	1.622

4 Conclusions

In this paper, we propose a novel symmetric flow model, named SyMOT-Flow, to learn a transformation between two unknown distributions from a set of samples drawn from each distribution, respectively. SyMOT-Flow leverages the maximum mean discrepancy (MMD) as a metric for comparing distributions. To enhance the generative performance of both forward and backward flows, a symmetrical design of the reversed component is incorporated based on the invertibility of the flow structure. Additionally, by treating the MMD as a relaxation of the equality constraint in the original optimal transport (OT) problem, SyMOT-Flow can also learn an asymptotic solution to OT. Besides, we provide some theoretical results regarding the feasibility of the proposed model and the connections to the OT solution. In the experimental section, SyMOT-Flow is evaluated on toy examples for illustration and real-world datasets, showcasing the generative samples and the transformation process achieved by the model. Furthermore, ablation studies are conducted to investigate the impact of the reversed flow constraint and the weight parameter on the OT cost. These experiments contribute to a better understanding of the effectiveness and robustness of SyMOT-Flow in practical scenarios.

References

[1] Antonio Criminisi, Jamie Shotton, Ender Konukoglu, et al. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and trends® in computer graphics and vision*, 7(2–3):81–227, 2012.

[2] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[3] Lars Ruthotto and Eldad Haber. An introduction to deep generative modeling. *GAMM-Mitteilungen*, 44(2):e202100008, 2021.

[4] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021.

[5] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[6] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020.

[7] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[8] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[9] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[10] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[11] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

[12] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

[13] Leonid Vitalevich Kantorovich. On a problem of Monge. *J. Math. Sci.(NY)*, 133:1383, 2006.

[14] SS Vallender. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.

[15] Ludger Rüschendorf. The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1):117–129, 1985.

[16] Cédric Villani and Cédric Villani. The wasserstein distances. *Optimal Transport: Old and New*, pages 93–111, 2009.

[17] Florentin Coeurdoux, Nicolas Dobigeon, and Pierre Chainais. Learning optimal transport between two empirical distributions with normalizing flows. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part V*, pages 275–290. Springer, 2023.

[18] Florentin Coeurdoux, Nicolas Dobigeon, and Pierre Chainais. Sliced-Wasserstein normalizing flows: beyond maximum likelihood training. *arXiv preprint arXiv:2207.05468*, 2022.

[19] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

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

[21] Biwei Dai and Uros Seljak. Sliced iterative normalizing flows. *arXiv preprint arXiv:2007.00674*, 2020.

[22] Samuel Klein, John Andrew Raine, and Tobias Golling. Flows for Flows: Training Normalizing Flows Between Arbitrary Distributions with Maximum Likelihood Estimation. *arXiv preprint arXiv:2211.02487*, 2022.

[23] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

[24] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.

[25] Tianyang Hu, Zixiang Chen, Hanxi Sun, Jincheng Bai, Mao Ye, and Guang Cheng. Stein neural sampler. *arXiv preprint arXiv:1810.03545*, 2018.

[26] Matthew Fisher, Tui Nolan, Matthew Graham, Dennis Prangle, and Chris Oates. Measure transport with kernel stein discrepancy. In *International Conference on Artificial Intelligence and Statistics*, pages 1054–1062. PMLR, 2021.

[27] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[28] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017.

[29] Wei Wang, Yuan Sun, and Saman Halgamuge. Improving MMD-GAN training with repulsive loss function. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HygjqjR9Km`.

[30] Zhifeng Kong and Kamalika Chaudhuri. Universal approximation of residual flows in maximum mean discrepancy. *arXiv preprint arXiv:2103.05793*, 2021.

[31] Sebastian Neumayer and Gabriele Steidl. From optimal transport to discrepancy. *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision*, pages 1–36, 2021.

[32] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[33] Carl-Johann Simon-Gabriel, Alessandro Barp, Bernhard Schölkopf, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies. *arXiv preprint arXiv:2006.09268*, 2020.

[34] Lynton Ardizzone, Till Bungert, Felix Draxler, Ullrich Köthe, Jakob Kruse, Robert Schmier, and Peter Sorrenson. Framework for Easily Invertible Architectures (FrEIA), 2018-2022. URL `https://github.com/vislearn/FrEIA`.

[35] Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. *Advances in neural information processing systems*, 25, 2012.