

APPENDIX A PROOF OF THE INJECTIVITY OF THE SPATIAL RADON TRANSFORM

We prove that the spatial Radon transform defined with a mapping $g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\theta}$ is injective if and only if $g(\cdot)$ is injective. In the following contents, we use $P_k(\mathbb{R}^d)$ to denote a set of Borel probability measures with finite k -th moment on \mathbb{R}^d , and $f_1 \equiv f_2$ is used to denote functions $f_1(\cdot) : X \rightarrow \mathbb{R}$ and $f_2(\cdot) : X \rightarrow \mathbb{R}$ that satisfy $f_1(x) = f_2(x)$ for $\forall x \in X$, and $f_1 \not\equiv f_2$ is used to denote functions $f_1(\cdot) : X \rightarrow \mathbb{R}$ and $f_2(\cdot) : X \rightarrow \mathbb{R}$ that satisfy $f_1(x) \neq f_2(x)$ for certain $x \in X$. With a slight abuse of notation, we interchangeably use $f_1(x) \equiv f_2(x)$ for $\forall x \in X$ and $f_1 \equiv f_2$.

Proof. By using proof by contradiction, we first prove that if $g(\cdot)$ is injective, the corresponding spatial Radon transform is injective. If the spatial Radon transform defined with an injective mapping $g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\theta}$ is not injective, there exist $\mu, \nu \in P_k(\mathbb{R}^d)$, $\mu \neq \nu$, such that $\mathcal{H}p_\mu(t, \theta; g) \equiv \mathcal{H}p_\nu(t, \theta; g)$ for $\forall t \in \mathbb{R}$ and $\forall \theta \in \mathbb{S}^{d_\theta-1}$, where p_μ and p_ν are probability density functions defined on \mathbb{R}^d and $p_\mu \neq p_\nu$.

From Equation (12), for $\forall t \in \mathbb{R}$ and $\forall \theta \in \mathbb{S}^{d_\theta-1}$, the spatial Radon transform can be written as:

$$\mathcal{H}p_\mu(t, \theta; g) = \mathcal{R}p_{\hat{\mu}_g}(t, \theta), \quad (20)$$

$$\mathcal{H}p_\nu(t, \theta; g) = \mathcal{R}p_{\hat{\nu}_g}(t, \theta), \quad (21)$$

where $p_{\hat{\mu}_g}$ and $p_{\hat{\nu}_g}$ refer to the probability density functions of $\hat{x} = g(x)$ and $\hat{y} = g(y)$ respectively, where $x \sim \mu$ and $y \sim \nu$. From Equations (20) and (21), we know $\mathcal{R}p_{\hat{\mu}_g}(t, \theta) \equiv \mathcal{R}p_{\hat{\nu}_g}(t, \theta)$ for $\forall t \in \mathbb{R}$ and $\forall \theta \in \mathbb{S}^{d_\theta-1}$, which implies $p_{\hat{\mu}_g} \equiv p_{\hat{\nu}_g}$ as the Radon transform is injective.

Since $g(\cdot)$ is injective, for $\forall \mathcal{X} \subseteq \mathbb{R}^d$, $x \in \mathcal{X}$ if and only if $\hat{x} = g(x) \in g(\mathcal{X})$, which implies $P(x \in \mathcal{X}) = P(\hat{x} \in g(\mathcal{X}))$, $P(y \in \mathcal{X}) = P(\hat{y} \in g(\mathcal{X}))$. Therefore,

$$\int_{g(\mathcal{X})} p_{\hat{\mu}_g}(\hat{x}) d\hat{x} = \int_{\mathcal{X}} p_\mu(x) dx, \quad (22)$$

$$\int_{g(\mathcal{X})} p_{\hat{\nu}_g}(\hat{y}) d\hat{y} = \int_{\mathcal{X}} p_\nu(y) dy. \quad (23)$$

Since $p_{\hat{\mu}_g} \equiv p_{\hat{\nu}_g}$, from Equations (22) and (23): $\int_{\mathcal{X}} p_\mu(x) dx = \int_{\mathcal{X}} p_\nu(y) dy$ for $\forall \mathcal{X} \subseteq \mathbb{R}^d$. Hence, for $\forall \mathcal{X} \subseteq \mathbb{R}^d$:

$$\int_{\mathcal{X}} (p_\mu(x) - p_\nu(x)) dx = 0, \quad (24)$$

which implies $p_\mu \equiv p_\nu$, contradicting with the assumption $p_\mu \neq p_\nu$. Therefore, if $\mathcal{H}p_\mu \equiv \mathcal{H}p_\nu$, $p_\mu \equiv p_\nu$. In addition, from the definition of the spatial Radon transform in Equation (11), it is trivial to show that if $p_\mu \equiv p_\nu$, $\mathcal{H}p_\mu(t, \theta; g) \equiv \mathcal{H}p_\nu(t, \theta; g)$. Therefore, $\mathcal{H}p_\mu \equiv \mathcal{H}p_\nu$ if and only if $p_\mu \equiv p_\nu$, i.e. the spatial Radon transform \mathcal{H} defined with an injective mapping $g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\theta}$ is injective.

We now prove that if the spatial Radon transform defined with a mapping $g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\theta}$ is injective, $g(\cdot)$ must be injective. Again, we use proof by contradiction. If $g(\cdot)$ is not injective, there exist $x_0, y_0 \in \mathbb{R}^d$ such that $x_0 \neq y_0$ and $g(x_0) = g(y_0)$. For two Dirac measures μ_1 and ν_1 which probability density functions are $p_{\mu_1}(x) = \delta(x - x_0)$ and $p_{\nu_1}(y) = \delta(y - y_0)$, respectively, we know $\mu_1 \neq \nu_1$ as $x_0 \neq y_0$.

We define variables $x \sim \mu_1$ and $y \sim \nu_1$. Then for variables $\hat{x} = g(x)$ and $\hat{y} = g(y)$, we denote their probability density functions by p_{μ_2} and p_{ν_2} , respectively. It is trivial to derive

$$p_{\mu_2}(\hat{x}) = \delta(\hat{x} - g(x_0)), \quad (25)$$

$$p_{\nu_2}(\hat{y}) = \delta(\hat{y} - g(y_0)), \quad (26)$$

which implies $p_{\mu_2} \equiv p_{\nu_2}$ as $g(x_0) = g(y_0)$.

From Equations (20), (21), (25) and (26), for $\forall t \in \mathbb{R}$ and $\forall \theta \in \mathbb{S}^{d_\theta-1}$:

$$\begin{aligned} \mathcal{H}p_{\mu_1}(t, \theta; g) &= \mathcal{R}p_{\mu_2}(t, \theta), \\ &= \mathcal{R}p_{\nu_2}(t, \theta), \\ &= \mathcal{H}p_{\nu_1}(t, \theta; g), \end{aligned} \quad (27)$$

which implies $\mathcal{H}p_{\mu_1} \equiv \mathcal{H}p_{\nu_1}$, contradicting with the assumption that the spatial Radon transform is injective. Therefore, if the spatial Radon transform is injective, $g(\cdot)$ must be injective. We conclude that the spatial Radon transform is injective if and only if the mapping $g(\cdot)$ is an injection. \square

APPENDIX B PROOF OF REMARK 2

We provide a proof for the claim in Remark 2 that the spatial Radon transform includes the vanilla Radon transform and the polynomial GRT as special cases.

Proof. Given a probability measure $\mu \in P(\mathbb{R}^d)$ which probability density function is p_μ , the spatial Radon transform of p_μ is defined as:

$$\mathcal{H}p_\mu(t, \theta; g) = \int_{\mathbb{R}^d} p_\mu(x) \delta(t - \langle g(x), \theta \rangle) dx, \quad (28)$$

where $t \in \mathbb{R}$ and $\theta \in \mathbb{S}^{d_\theta - 1}$ are the parameters of hypersurfaces in \mathbb{R}^d . When the mapping $g(\cdot)$ is an identity mapping, i.e. $g(x) = x$ for $\forall x \in \mathbb{R}^d$, the spatial Radon transform degenerates to the vanilla Radon transform:

$$\begin{aligned} \mathcal{H}p_\mu(t, \theta; g) &= \int_{\mathbb{R}^d} p_\mu(x) \delta(t - \langle x, \theta \rangle) dx \\ &= \mathcal{R}p_\mu(t, \theta). \end{aligned} \quad (29)$$

Ehrenpreis (2003) provides a class of injective GRTs named polynomial GRTs by adopting homogeneous polynomial functions with an odd degree m as the defining function:

$$\begin{aligned} \mathcal{G}p_\mu(t, \theta) &= \int_{\mathbb{R}^d} p_\mu(x) \delta(t - \sum_{i=1}^{d_\alpha} \theta_i x^{\alpha_i}) dx, \\ &\text{s.t. } |\alpha_i| = m, \end{aligned} \quad (30)$$

where $\alpha_i = (\eta_{i,1}, \dots, \eta_{i,d}) \in \mathbb{N}^d$, $|\alpha_i| = \sum_{j=1}^d \eta_{i,j}$, $x^{\alpha_i} = \prod_{j=1}^d x_j^{\eta_{i,j}}$ for $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, d_α is the number of all possible multi-indices α_i that satisfies $|\alpha_i| = m$, and $\theta = (\theta_1, \dots, \theta_{d_\alpha}) \in \mathbb{S}^{d_\alpha - 1}$.

In spatial Radon transform, for $\forall x \in \mathbb{R}^d$, when the mapping $g(\cdot)$ is defined as:

$$g(x) = (x^{\alpha_1}, \dots, x^{\alpha_{d_\alpha}}), \quad (31)$$

the spatial Radon transform is equivalent to the polynomial GRT defined in Equation (30):

$$\begin{aligned} \mathcal{H}p_\mu(t, \theta; g) &= \int_{\mathbb{R}^d} p_\mu(x) \delta(t - \langle g(x), \theta \rangle) dx \\ &= \int_{\mathbb{R}^d} p_\mu(x) \delta(t - \sum_{i=1}^{d_\alpha} \theta_i x^{\alpha_i}) dx. \end{aligned} \quad (32)$$

□

APPENDIX C PROOF OF THEOREM 1

We provide a proof that the ASWD defined with a mapping $g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\theta}$ is a metric on $P_k(\mathbb{R}^d)$, if and only if $g(\cdot)$ is injective. In what follows, we denote a set of Borel probability measures with finite k -th moment on \mathbb{R}^d by $P_k(\mathbb{R}^d)$, and use $\mu, \nu \in P_k(\mathbb{R}^d)$ to refer to two probability measures whose probability density functions are p_μ and p_ν .

Proof. Symmetry: Since the k -Wasserstein distance is a metric thus symmetric (Villani, 2008):

$$W_k(\mathcal{H}p_\mu(\cdot, \theta; g), \mathcal{H}p_\nu(\cdot, \theta; g)) = W_k(\mathcal{H}p_\nu(\cdot, \theta; g), \mathcal{H}p_\mu(\cdot, \theta; g)). \quad (33)$$

Therefore,

$$\begin{aligned} \text{ASWD}_k(\mu, \nu; g) &= \left(\int_{\mathbb{S}^{d_\theta - 1}} W_k^k(\mathcal{H}p_\mu(\cdot, \theta; g), \mathcal{H}p_\nu(\cdot, \theta; g)) d\theta \right)^{\frac{1}{k}} \\ &= \left(\int_{\mathbb{S}^{d_\theta - 1}} W_k^k(\mathcal{H}p_\nu(\cdot, \theta; g), \mathcal{H}p_\mu(\cdot, \theta; g)) d\theta \right)^{\frac{1}{k}} \\ &= \text{ASWD}_k(\nu, \mu; g). \end{aligned} \quad (34)$$

Triangle inequality: Given an injective mapping $g(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\theta}$ and probability measures $\mu_1, \mu_2, \mu_3 \in P_k(\mathbb{R}^d)$, since the k -Wasserstein distance satisfies the triangle inequality (Villani, 2008), the following inequality holds:

$$\begin{aligned} \text{ASWD}_k(\mu_1, \mu_3; g) &= \left(\int_{\mathbb{S}^{d_\theta-1}} W_k^k(\mathcal{H}p_{\mu_1}(\cdot, \theta; g), \mathcal{H}p_{\mu_3}(\cdot, \theta; g)) d\theta \right)^{\frac{1}{k}} \\ &\leq \left(\int_{\mathbb{S}^{d_\theta-1}} (W_k(\mathcal{H}p_{\mu_1}(\cdot, \theta; g), \mathcal{H}p_{\mu_2}(\cdot, \theta; g)) + W_k(\mathcal{H}p_{\mu_2}(\cdot, \theta; g), \mathcal{H}p_{\mu_3}(\cdot, \theta; g)))^k d\theta \right)^{\frac{1}{k}} \\ &\leq \left(\int_{\mathbb{S}^{d_\theta-1}} W_k^k(\mathcal{H}p_{\mu_1}(\cdot, \theta; g), \mathcal{H}p_{\mu_2}(\cdot, \theta; g)) d\theta \right)^{\frac{1}{k}} + \left(\int_{\mathbb{S}^{d_\theta-1}} W_k^k(\mathcal{H}p_{\mu_2}(\cdot, \theta; g), \mathcal{H}p_{\mu_3}(\cdot, \theta; g)) d\theta \right)^{\frac{1}{k}} \\ &= \text{ASWD}_k(\mu_1, \mu_2; g) + \text{ASWD}_k(\mu_2, \mu_3; g), \end{aligned}$$

where the second inequality is due to the Minkowski inequality in $L^k(\mathbb{S}^{d_\theta-1})$.

Identity of indiscernibles: Since $W_k(\mu, \mu) = 0$ for $\forall \mu \in P_k(\mathbb{R}^d)$, we have

$$\text{ASWD}_k(\mu, \mu; g) = \left(\int_{\mathbb{S}^{d_\theta-1}} W_k^k(\mathcal{H}p_\mu(\cdot, \theta; g), \mathcal{H}p_\mu(\cdot, \theta; g)) d\theta \right)^{\frac{1}{k}} = 0, \quad (35)$$

for $\forall \mu \in P_k(\mathbb{R}^d)$.

Conversely, for $\forall \mu, \nu \in P_k(\mathbb{R}^d)$, if $\text{ASWD}_k(\mu, \nu; g) = 0$, from the definition of the ASWD:

$$\text{ASWD}_k(\mu, \nu; g) = \left(\int_{\mathbb{S}^{d_\theta-1}} W_k^k(\mathcal{H}p_\mu(\cdot, \theta; g), \mathcal{H}p_\nu(\cdot, \theta; g)) d\theta \right)^{\frac{1}{k}} = 0, \quad (36)$$

which implies $W_k(\mathcal{H}p_\mu(\cdot, \theta; g), \mathcal{H}p_\nu(\cdot, \theta; g)) = 0$ for $\forall \theta \in \mathbb{S}^{d_\theta-1}$. Due to the non-negativity of k -th Wasserstein distance as it is a metric on $P_k(\mathbb{R}^d)$, $W_k(\mathcal{H}p_\mu(\cdot, \theta; g), \mathcal{H}p_\nu(\cdot, \theta; g)) = 0$ holds for $\forall \theta \in \mathbb{S}^{d_\theta-1}$ if and only if $\mathcal{H}p_\mu(\cdot, \theta; g) \equiv \mathcal{H}p_\nu(\cdot, \theta; g)$. Again, given the spatial Radon transform is injective when $g(\cdot)$ is injective (see the proof in Appendix A), $\mathcal{H}p_\mu(\cdot, \theta; g) \equiv \mathcal{H}p_\nu(\cdot, \theta; g)$ implies $p_\mu \equiv p_\nu$ and $\mu \equiv \nu$ if $g(\cdot)$ is injective.

In addition, if $g(\cdot)$ is not injective, the spatial Radon transform is not injective (see the proof in Appendix A), then $\exists \mu, \nu \in P_k(\mathbb{R}^d)$, $\mu \not\equiv \nu$ such that $\mathcal{H}p_\mu(\cdot, \theta; g) \equiv \mathcal{H}p_\nu(\cdot, \theta; g)$, which implies $\text{ASWD}_k(\mu, \nu; g) = 0$ for $\mu \not\equiv \nu$. Therefore, the ASWD satisfies the identity of indiscernibles if and only if $g(\cdot)$ is injective.

Non-negativity: The three axioms of a distance metric, i.e. symmetry, triangle inequality, and identity of indiscernibles imply the non-negativity of the ASWD. Since the Wasserstein distance is non-negative, for $\forall \mu, \nu \in P_k(\mathbb{R}^d)$, it can also be straightforwardly proved the ASWD between μ and ν is non-negative:

$$\begin{aligned} \text{ASWD}_k(\mu, \nu; g) &= \left(\int_{\mathbb{S}^{d_\theta-1}} W_k^k(\mathcal{H}p_\mu(\cdot, \theta; g), \mathcal{H}p_\nu(\cdot, \theta; g)) d\theta \right)^{\frac{1}{k}} \\ &\geq \left(\int_{\mathbb{S}^{d_\theta-1}} 0^k d\theta \right)^{\frac{1}{k}} = 0. \end{aligned} \quad (37)$$

Therefore, the ASWD is a metric on $P_k(\mathbb{R}^d)$ if and only if $g(\cdot)$ is injective. \square

APPENDIX D PSEUDOCODE FOR THE EMPIRICAL VERSION OF THE ASWD

Algorithm 1 The augmented sliced Wasserstein distance. All of the for loops can be parallelized.

Require: Sets of samples $\{x_n \in \mathbb{R}^d\}_{n=1}^N, \{y_n \in \mathbb{R}^d\}_{n=1}^N$;

Require: Randomly initialized injective neural network $g_\omega(\cdot): \mathbb{R}^d \rightarrow \mathbb{R}^{d_\theta}$;

Require: Number of projections L , hyperparameter λ , learning rate ϵ , number of iterations M ;

- 1: Initialize $D=0, L_\lambda=0, m=1$;
 - 2: **while** ω has not converged and $m \leq M$ **do**
 - 3: Draw a set of samples $\{\theta_l\}_{l=1}^L$ from $\mathbb{S}^{d_\theta-1}$;
 - 4: **for** $n=1$ to N **do**
 - 5: Compute $g_\omega(x_n)$ and $g_\omega(y_n)$;
 - 6: Calculate the regularization term $L_\lambda \leftarrow L_\lambda + \frac{\lambda}{N} (\|g_\omega(x_n)\|_2 + \|g_\omega(y_n)\|_2)$;
 - 7: **end for**
 - 8: **for** $l=1$ to L **do**
 - 9: Compute $\beta(x_n, \theta_l) = \langle g_\omega(x_n), \theta_l \rangle, \beta(y_n, \theta_l) = \langle g_\omega(y_n), \theta_l \rangle$ for each n ;
 - 10: Sort $\beta(x_n, \theta_l)$ and $\beta(y_n, \theta_l)$ in ascending order s.t. $\beta(x_{I_x^l[n]}, \theta_l) \leq \beta(x_{I_x^l[n+1]}, \theta_l)$ and $\beta(y_{I_y^l[n]}, \theta_l) \leq \beta(y_{I_y^l[n+1]}, \theta_l)$;
 - 11: Calculate the ASWD: $D \leftarrow D + (\frac{1}{L} \sum_{n=1}^N |\beta(x_{I_x^l[n]}, \theta_l) - \beta(y_{I_y^l[n]}, \theta_l)|^k)^{\frac{1}{k}}$;
 - 12: **end for**
 - 13: $\mathcal{L} \leftarrow D - L_\lambda$;
 - 14: Update ω by gradient ascent $\omega \leftarrow \omega + \epsilon \cdot \nabla_\omega \mathcal{L}$;
 - 15: Reset $D=0, L_\lambda=0$, update $m \leftarrow m+1$;
 - 16: **end while**
 - 17: Draw a set of samples $\{\theta_l\}_{l=1}^L$ from $\mathbb{S}^{d_\theta-1}$;
 - 18: **for** $n=1$ to N **do**
 - 19: Compute $g_\omega(x_n)$ and $g_\omega(y_n)$;
 - 20: **end for**
 - 21: **for** $l=1$ to L **do**
 - 22: Compute $\beta(x_n, \theta_l) = \langle g_\omega(x_n), \theta_l \rangle, \beta(y_n, \theta_l) = \langle g_\omega(y_n), \theta_l \rangle$ for each n ;
 - 23: Sort $\beta(x_n, \theta_l)$ and $\beta(y_n, \theta_l)$ in ascending order s.t. $\beta(x_{I_x^l[n]}, \theta_l) \leq \beta(x_{I_x^l[n+1]}, \theta_l)$ and $\beta(y_{I_y^l[n]}, \theta_l) \leq \beta(y_{I_y^l[n+1]}, \theta_l)$;
 - 24: Calculate the ASWD: $D \leftarrow D + (\frac{1}{L} \sum_{n=1}^N |\beta(x_{I_x^l[n]}, \theta_l) - \beta(y_{I_y^l[n]}, \theta_l)|^k)^{\frac{1}{k}}$;
 - 25: **end for**
 - 26: **Output:** Augmented sliced Wasserstein distance D .
-

APPENDIX E EXPERIMENTAL SETUPS

E.1 HYPERPARAMETERS IN THE SLICED WASSERSTEIN FLOW EXPERIMENT

We randomly generate 500 samples both for target distributions and source distributions. We initialize the source distributions μ_0 as standard normal distributions $\mathcal{N}(0, I)$, where I is a 2-dimensional identity matrix. We update source distributions using Adam optimizer (Kingma & Ba, 2015), and set the learning rate=0.002. For all methods, we set the order $k=2$. When testing the ASWD, the number of iterations M in Algorithm 1 is set to 10. Empirical errors in the experiment are found to be not sensitive to the choice of λ in a candidate set of {0.01, 0.05, 0.1, 0.5}. The reported results are produced with $\lambda=0.1$.

E.2 NETWORK ARCHITECTURE IN THE GENERATIVE MODELING EXPERIMENT

Denote a convolutional layer whose kernel size is s with C kernels by $Conv_C(s \times s)$, and a fully-connected layer whose input and output layer have s_1 and s_2 neurons by $FC(s_1 \times s_2)$. The network structure used in the generative modeling experiment is configured to be the same as described in (Nguyen et al., 2020):

$$\begin{aligned}
 h_\psi &: (64 \times 64 \times 3) \rightarrow Conv_{64}(4 \times 4) \rightarrow LeakyReLU(0.2) \rightarrow Conv_{128}(4 \times 4) \rightarrow BatchNormalization \\
 &\rightarrow LeakyReLU(0.2) \rightarrow Conv_{256}(4 \times 4) \rightarrow BatchNormalization \rightarrow LeakyReLU(0.2) \rightarrow \\
 &Conv_{512}(4 \times 4) \rightarrow BatchNormalization \rightarrow Tanh \xrightarrow{Output} (512 \times 4 \times 4) \\
 D_\Psi &: Conv_1(4 \times 4) \rightarrow Sigmoid \xrightarrow{Output} (1 \times 1 \times 1) \\
 G_\Phi &: z \in \mathbb{R}^{32} \rightarrow ConvTranspose_{512}(4 \times 4) \rightarrow BatchNormalization \rightarrow ReLU \rightarrow \\
 &ConvTranspose_{256}(4 \times 4) \rightarrow BatchNormalization \rightarrow ReLU \rightarrow ConvTranspose_{128}(4 \times 4) \rightarrow \\
 &BatchNormalization \rightarrow ReLU \rightarrow ConvTranspose_{64}(4 \times 4) \rightarrow BatchNormalization \rightarrow \\
 &ConvTranspose_3(4 \times 4) \rightarrow Tanh \xrightarrow{Output} (64 \times 64 \times 3) \\
 \phi &: FC(8192 \times 8192) \xrightarrow{Output} (8192)\text{-dimensional vector}
 \end{aligned}$$

We train the models with the Adam optimizer (Kingma & Ba, 2015), and set the batch size to 512. Following the setup in (Nguyen et al., 2020), the learning rate is set to 0.0005 and beta=(0.5, 0.999) for both CIFAR10 dataset and CELEBA dataset. For all methods, we set the order k to 2. For the ASWD, the number of iterations M in Algorithm 1 is set to 5. The hyperparameter λ is set to 0.5 to introduce slightly larger regularization of the optimization objective due to the small output values from the feature layer h_ψ .

APPENDIX F ADDITIONAL RESULTS IN THE SLICED WASSERSTEIN FLOW EXPERIMENT

F.1 FULL EXPERIMENTAL RESULTS ON THE SLICED WASSERSTEIN EXPERIMENT

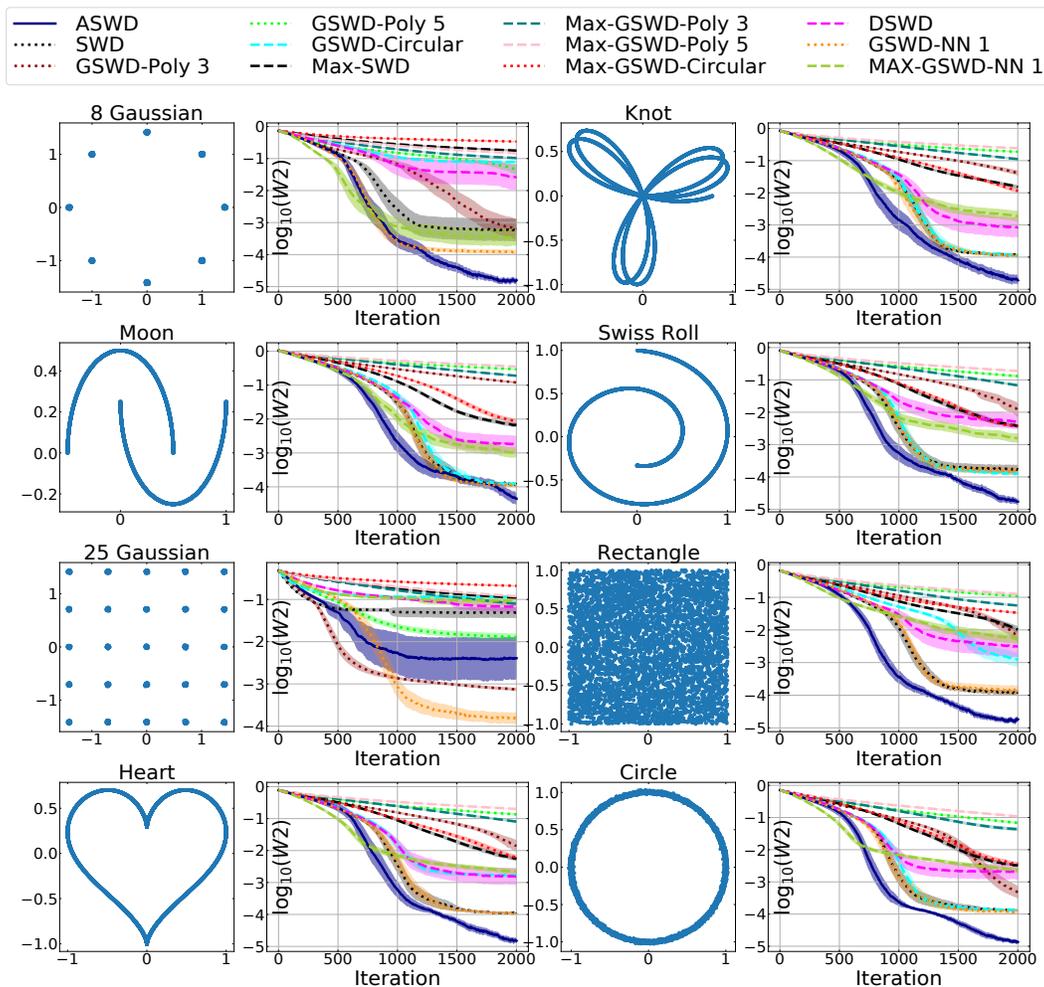


Figure 4: Full experimental results on the sliced Wasserstein flow example. The first and third columns are target distributions. The second and fourth columns are log 2-Wasserstein distances between the target distributions and the source distributions. The horizontal axis shows the number of training iterations. Solid lines and shaded areas represent the average values and 95% confidence intervals of log 2-Wasserstein distances over 50 runs.

F.2 ABLATION STUDY

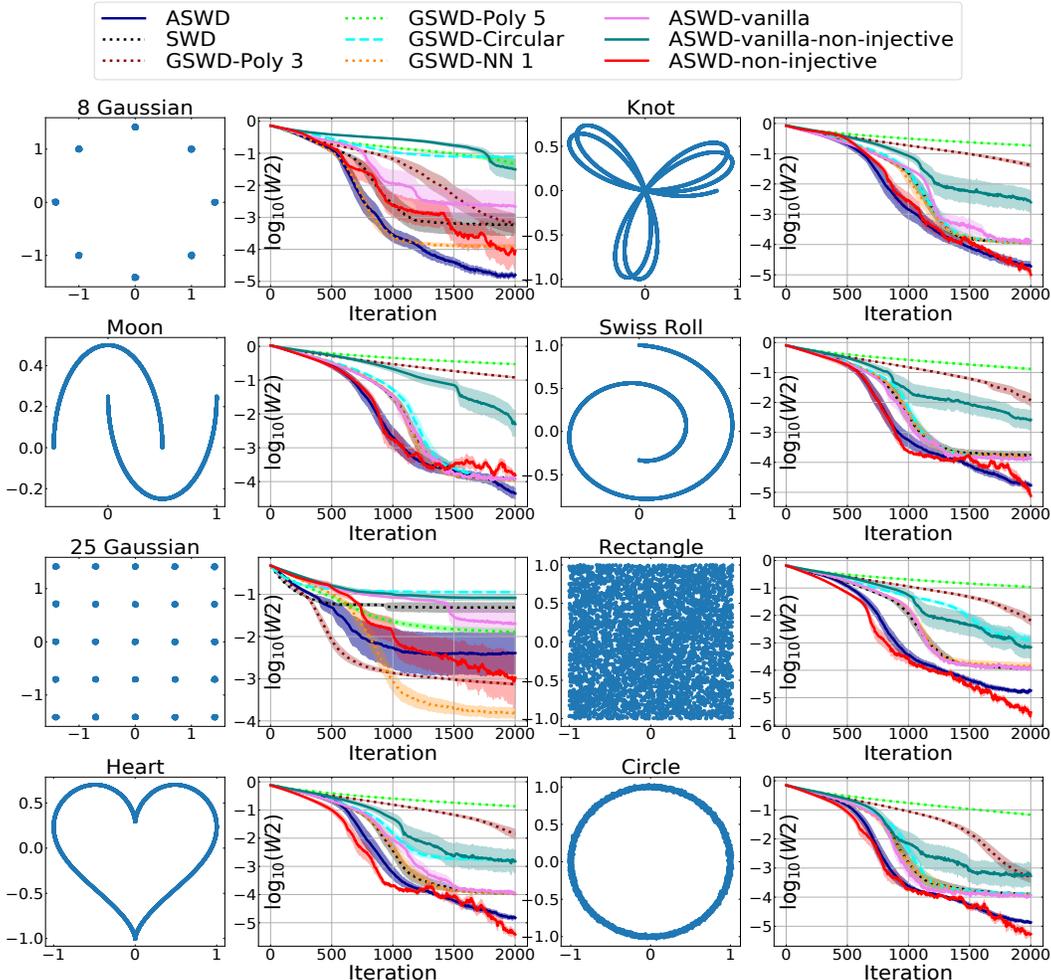


Figure 5: Ablation study on the impact from injective neural networks and the optimization of hypersurfaces on the ASWD. ASWDs with different mappings are compared to GSWDs with different defining functions. The first and third columns show target distributions. The second and fourth columns plot \log_{10} 2-Wasserstein distances between the target distributions and the source distributions. In the second and fourth columns, the horizontal axis shows the number of training iterations. Solid lines and shaded areas represent the average values and 95% confidence intervals of \log_{10} 2-Wasserstein distances over 50 runs.

Impact of the injectivity of the mapping

In this ablation study, we first compare ASWDs constructed by different mappings to GSWDs with different predefined defining functions, and investigate the effects of the optimization and injectivity of the adopted mapping $g_\omega(\cdot)$ used in the ASWDs. In what follows, “ASWD-vanilla” is used to denote ASWDs that employ randomly initialized neural network $\phi_\omega(\cdot)$ to parameterize the injective mapping $g_\omega(\cdot) = [\cdot, \phi_\omega(\cdot)]$, i.e. the mapping $g_\omega(\cdot)$ is not optimized in the ASWD-vanilla and the results of ASWD-vanilla reported in Figure 5 are obtained by projecting samples onto random hypersurfaces. Furthermore, the “ASWD-non-injective” refers to ASWDs that do not use the injectivity trick, i.e. the mapping $g_\omega(\cdot) = \phi_\omega(\cdot)$ is not guaranteed to be injective. In addition, the “ASWD-vanilla-non-injective” adopts both setups in the “ASWD-vanilla” and “ASWD-non-injective”, resulting in a random non-injective mapping $g_\omega(\cdot)$. The reported experiment results in this ablation study is calculated over 50 runs, and the neural network $\phi_\omega(\cdot)$ is reinitialized randomly in each run.

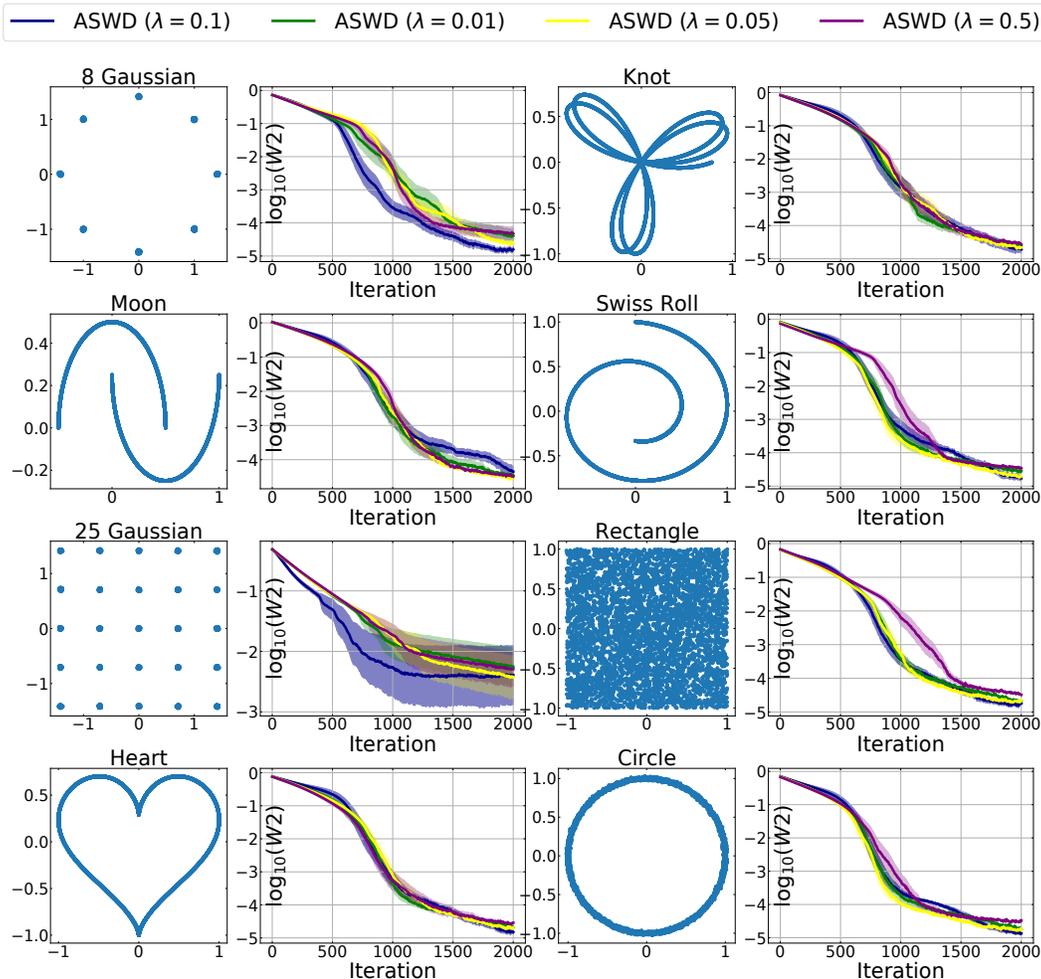


Figure 6: Ablation study on the impact from the value of λ . The performance of the ASWD with different values of λ are compared. The first and third columns show target distributions. The second and fourth columns plot $\log 2$ -Wasserstein distances between the target distributions and the source distributions. In the second and fourth columns, the horizontal axis shows the number of training iterations. Solid lines and shaded areas represent the average values and 95% confidence intervals of $\log 2$ -Wasserstein distances over 50 runs.

From Figure 5, it can be observed that the ASWD-vanilla shows comparable performance to GSWDs defined by polynomial and circular defining functions, which implies GSWDs with predefined defining functions are as uninformative as projecting distributions onto random hypersurfaces constructed by the ASWD. In GSWDs, the hypersurfaces are predefined and cannot be optimized since they are determined by the functional forms of the defining functions. On the contrary, we found that the optimization of hypersurfaces in the ASWD framework can help improve the performance of the slice-based Wasserstein distance. As in Figure 5, the ASWD and the ASWD-non-injective present significantly better performance than methods that do not optimize their hypersurfaces (ASWD-vanilla, ASWD-vanilla-non-injective, and GSWDs). In terms of the impact of the injectivity of the mapping g_ω , in this experiment, the ASWD-vanilla exhibits lower 2-Wasserstein distances than the ASWD-vanilla-non-injective in all tested distributions, and the ASWD leads to more stable training than the ASWD-non-injective. Therefore, the injectivity of the mapping $g_\omega(\cdot)$ does not only guarantee the ASWD to be a valid distance metric as proved in Section 3, but also better empirical performance in this experiment setup.

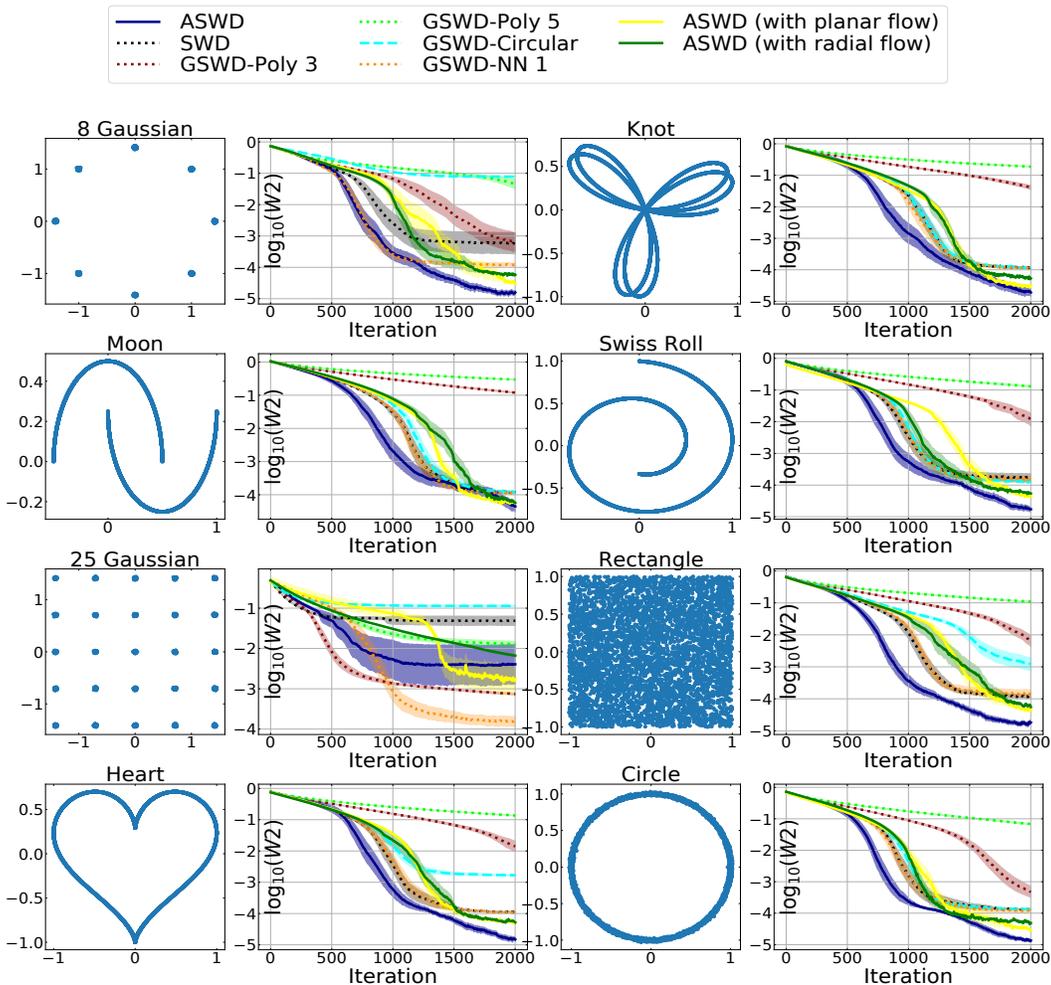


Figure 7: Ablation study on the impact from the choice of injective networks. The performance of the ASWD with different types of injective networks are compared. The first and third columns show target distributions. The second and fourth columns plot \log_2 -Wasserstein distances between the target distributions and the source distributions. In the second and fourth columns, the horizontal axis shows the number of training iterations. Solid lines and shaded areas represent the average values and 95% confidence intervals of \log_2 -Wasserstein distances over 50 runs.

Impact of the regularization coefficient

In addition, the sensitivity of the ASWD to the value of λ in Equation (16) is also tested in this ablation study. In this experiment, the value of λ is selected from a candidate set $\{0.01, 0.05, 0.1, 0.5\}$. The numerical results in Figure 6 indicate that the performance of ASWD is not sensitive to the value of λ in that candidate set, as it can be observed that different values of λ lead to similar performance of the ASWD.

Choice of injective mapping

We reported in Figure 7 the performance of the ASWD defined with other types of injective mappings other than Equation (15). In particular, we examined two invertible mappings, including the planar flow and radial flow (Rezende & Mohamed, 2015), as alternatives to the injective mapping defined by Equation (15). The numerical results presented in Figure 7 show that the ASWD defined with planar flow and radial flow produced better performance than GSWD variants in most setups in Figure 7. They exhibit slightly worse performance compared with the ASWD with injective mapping defined in Equation (15), possibly due to the additional restriction in invertible mapping imposed the planar flow and radial flow.

APPENDIX G ADDITIONAL
RESULTS IN THE GENERATIVE MODELING EXPERIMENT

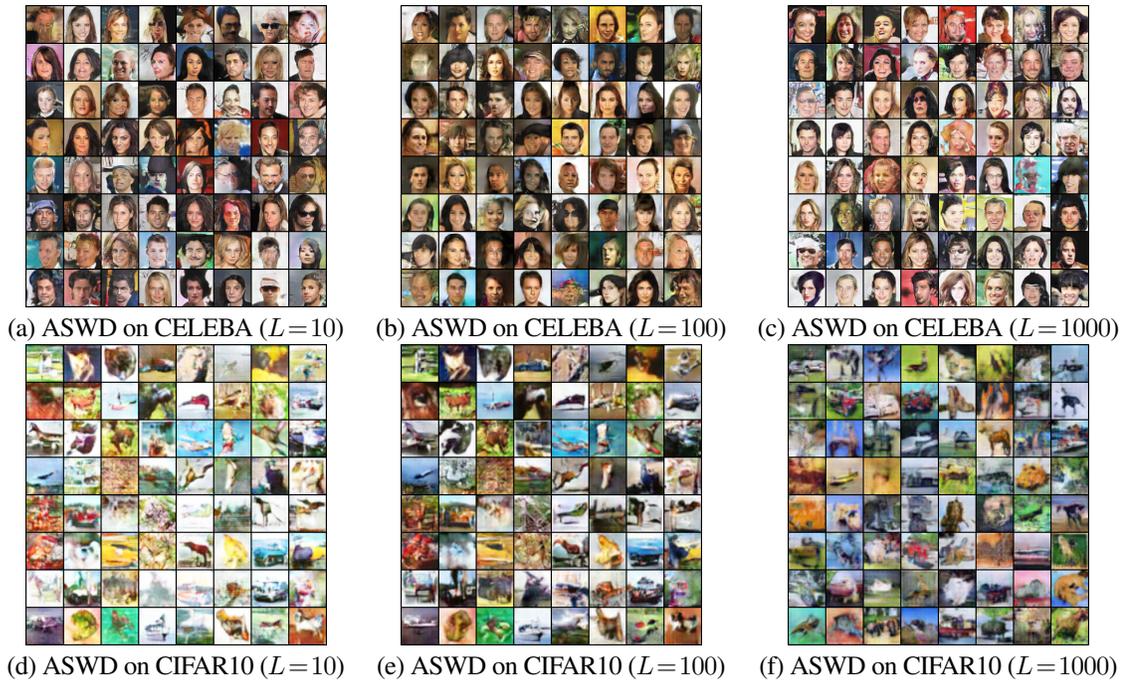


Figure 8: Visualized experimental results of the ASWD on CELEBA and CIFAR10 datasets with 10, 100, 1000 projections. The first row shows randomly selected samples of generated CELEBA images, the second row shows randomly selected samples of generated CIFAR10 images.



Figure 9: Visualized experimental results of the SWD, GSWD, and DSWD on CELEBA and CIFAR10 datasets with 1000 projections. The first row shows randomly selected samples of generated CELEBA images, the second row shows randomly selected samples of generated CIFAR10 images.