Supplementary Material: Lipschitz Continuous Autoencoders in Application to Anomaly Detection

In this supplementary material, we provide proofs for Propositions, Theorems, and Lemma in the manuscript in Appendix A, implementation details in Appendix B. Results of ablation study on KDD99 are provided in Appendix C, and experiment results on MNIST and Fashion-MNIST are provided in Appendix D.

A Proofs

Proof of Proposition 1. By definition of $T_c(\cdot, h)$, $T_c(\nu\delta_x + (1-\nu)\mathbb{P}_X^{(0)}, h)$ is equal to $\nu T_c(\delta_x, h) + (1-\nu)T_c(\mathbb{P}_X^{(0)}, h)$ for all $\nu \in (0, 1]$ and $h \in \mathcal{H}$. By rearranging terms, we conclude the proof. □

Proof of Proposition 2. Since T is the expected negative log-likelihood and $\mathcal{H}$ is the set of all probability density functions defined on $\mathcal{X}$, the objective function for a distribution $\mathbb{P}$ can be expressed as $T_{\mathrm{L}}(\mathbb{P}, h) = \mathcal{D}_{\mathrm{KL}}(\mathbb{P}||H) + S(\mathbb{P})$, where H is a distribution function associated to h. Thus, the optimizer for normal data is $h^{(0)} = d\mathbb{P}_X^{(0)}/dx$. The assumption $S(\mathbb{P}_X^{(1)}) \geq S(\mathbb{P}_X^{(0)})$ can be expressed as

$$\int_{\mathcal{X}} \log(d\mathbb{P}_X^{(1)}(x)/dx) d\mathbb{P}_X^{(1)}(x) \leq \int_{\mathcal{X}} \log(d\mathbb{P}_X^{(0)}(x)/dx) d\mathbb{P}_X^{(0)}(x), \tag{1}$$

and the assumption $\mathcal{D}_{\mathrm{KL}}(\mathbb{P}_X^{(1)}||\mathbb{P}_X^{(\nu)}) > \mathcal{D}_{\mathrm{KL}}(\mathbb{P}_X^{(0)}||\mathbb{P}_X^{(\nu)})$ can be expressed as

$$\int_{\mathcal{X}} \log \frac{d\mathbb{P}_X^{(0)}(x)/dx}{d\mathbb{P}_X^{(\nu)}(x)/dx} d\mathbb{P}_X^{(0)}(x) < \int_{\mathcal{X}} \log \frac{d\mathbb{P}_X^{(1)}(x)/dx}{d\mathbb{P}_X^{(\nu)}(x)/dx} d\mathbb{P}_X^{(1)}(x). \tag{2}$$

Then, adding (1) and (2) gives $\int_{\mathcal{X}} -\log(d\mathbb{P}_X^{(\nu)}(x)/dx) d\mathbb{P}_X^{(0)}(x) < \int_{\mathcal{X}} -\log(d\mathbb{P}_X^{(\nu)}(x)/dx) d\mathbb{P}_X^{(1)}(x)$, which is equivalent to $\int_{\mathcal{X}} -\log(d\mathbb{P}_X^{(\nu)}(x)/dx) d\mathbb{P}_X^{(0)}(x) < \int_{\mathcal{X}} -\log(d\mathbb{P}_X^{(\nu)}(x)/dx) d\mathbb{P}_X^{(\nu)}(x)$ since $d\mathbb{P}_X^{(\nu)}(x) = (1-\nu) d\mathbb{P}_X^{(0)}(x) + \nu d\mathbb{P}_X^{(1)}(x)$. Thus, $T_{\mathrm{L}}(\mathbb{P}_X^{(0)}, h^{(\nu)}) < T_{\mathrm{L}}(\mathbb{P}_X^{(\nu)}, h^{(\nu)})$ where $h^{(\nu)} := d\mathbb{P}_X^{(\nu)}/dx \in \arg\min_{h\in\mathcal{H}} T_{\mathrm{L}}(\mathbb{P}_X^{(\nu)}, h)$. By definition of $h^{(0)}$ and $h^{(\nu)}$, we have $T_{\mathrm{L}}(\mathbb{P}_X^{(0)}, h^{(0)}) \leq T_{\mathrm{L}}(\mathbb{P}_X^{(0)}, h^{(\nu)})$ and $T_{\mathrm{L}}(\mathbb{P}_X^{(\nu)}, h^{(\nu)}) \leq T_{\mathrm{L}}(\mathbb{P}_X^{(\nu)}, h^{(0)})$, which concludes the proof. □

Proof of Proposition 3. By definition of $T_c(\cdot, h)$, we have $T_c(\nu\mathbb{P}_X^{(1)} + (1-\nu)\mathbb{P}_X^{(0)}, h) = \nu T_c(\mathbb{P}_X^{(1)}, h) + (1-\nu)T_c(\mathbb{P}_X^{(0)}, h)$ for all $\nu \in (0, 1]$ and $h \in \mathcal{H}$. Hence, it implies $T_c(\nu\mathbb{P}_X^{(1)} + (1-\nu)\mathbb{P}_X^{(0)}, h) - T_c(\mathbb{P}_X^{(0)}, h) = \nu(T_c(\mathbb{P}_X^{(1)}, h) - T_c(\mathbb{P}_X^{(0)}, h))$ and the positiveness of ν concludes the proof. □

Proof of Example 1. For any $h \in \mathcal{H}$, the hidden layer has one node, so the output for input x can be expressed as $Ax + b$ where A is 2×2 matrix with rank 1 and $b \in \mathbb{R}^2$. Since the rank of A is 1, A can be expressed as $A = uv^T$ for some $u, v \in \mathbb{R}^2$. Since $X^{(0)}$ has mean and variance of 0_2 and I_2, respectively, $\mathbb{E}_{X\sim\mathbb{P}_X^{(0)}}(||X - (uv^T X + b)||_2^2) = 1 + ((v^T v)(u^T u) - 2v^T u + 1) + b^T b \geq 1$. For any a such that $a^T a = 1$, an autoencoder $h_a(x) := aa^T x$ is in $\mathcal{H}$ and has the reconstruction error of 1, so $T_c(\mathbb{P}_X^{(0)}, h_a) \leq T_c(\mathbb{P}_X^{(0)}, h)$ for all $h \in \mathcal{H}$. However, the reconstruction error of h_a on $\mathbb{P}_X^{(1)}$, $T_c(\mathbb{P}_X^{(1)}, h_a)$ is equal to $1 + \mu^T(I_2 - aa^T)\mu$, so the anomaly detection algorithm equipped with T_c and $\mathcal{H}$ is not admissible when μ is proportional to a. Since a can be any vector satisfying $a^T a = 1$, it concludes the proof. □

Lemma 1. *Let $\mathbb{P}, \mathbb{Q}$ be two probability measures defined on $\mathcal{X}$. For $0 \leq \beta \leq \alpha \leq 1$, set convex combinations of the two probability measures, $\alpha\mathbb{P} + (1-\alpha)\mathbb{Q}$ and $\beta\mathbb{P} + (1-\beta)\mathbb{Q}$. Then,*

$$\gamma_{\mathcal{F}}(\alpha\mathbb{P} + (1-\alpha)\mathbb{Q}, \beta\mathbb{P} + (1-\beta)\mathbb{Q}) = (\alpha - \beta)\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}).$$

Proof. Note that the definition of IPM is given by

$$\gamma_{\mathcal{F}}(\alpha\mathbb{P} + (1-\alpha)\mathbb{Q}, \beta\mathbb{P} + (1-\beta)\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left| \mathbb{E}_{X\sim\alpha\mathbb{P}+(1-\alpha)\mathbb{Q}}[f(X)] - \mathbb{E}_{X\sim\beta\mathbb{P}+(1-\beta)\mathbb{Q}}[f(X)] \right|.$$

We have

$$\begin{aligned}
&\mathbb{E}_{X\sim\alpha\mathbb{P}+(1-\alpha)\mathbb{Q}}[f(X)] - \mathbb{E}_{X\sim\beta\mathbb{P}+(1-\beta)\mathbb{Q}}[f(X)] \\
&= \{\alpha\mathbb{E}_{X\sim\mathbb{P}}[f(X)] + (1-\alpha)\mathbb{E}_{X\sim\mathbb{Q}}[f(X)]\} - \{\beta\mathbb{E}_{X\sim\mathbb{P}}[f(X)] + (1-\beta)\mathbb{E}_{X\sim\mathbb{Q}}[f(X)]\} \\
&= (\alpha - \beta)\{\mathbb{E}_{X\sim\mathbb{P}}[f(X)] - \mathbb{E}_{X\sim\mathbb{Q}}[f(X)]\}.
\end{aligned}$$

Hence,

$$\begin{aligned}\gamma_{\mathcal{F}}(\alpha\mathbb{P}+(1-\alpha)\mathbb{Q},\beta\mathbb{P}+(1-\beta)\mathbb{Q}) &= (\alpha-\beta)\sup_{f\in\mathcal{F}}\left|\mathbb{E}_{X\sim\mathbb{P}}[f(X)]-\mathbb{E}_{X\sim\mathbb{Q}}[f(X)]\right| \\ &= (\alpha-\beta)\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q}).\end{aligned}$$

This concludes the proof. □

Proof of Theorem 1. Since $\gamma_{\mathcal{F}}$ is a metric on $\Pi_{\mathcal{X}}$, it obeys a triangle inequality. Hence, for all $g \in \mathcal{H}_{\mathcal{F}}^{(K)}$, we have

$$\begin{aligned}\gamma_{\mathcal{F}}(\mathbb{P}_X^{(\nu)}, g\#\mathbb{P}_X^{(\nu)}) &\geq \gamma_{\mathcal{F}}(\mathbb{P}_X^{(\nu)},\mathbb{P}_X^{(0)}) - \gamma_{\mathcal{F}}(g\#\mathbb{P}_X^{(\nu)}, g\#\mathbb{P}_X^{(0)}) - \gamma_{\mathcal{F}}(g\#\mathbb{P}_X^{(0)},\mathbb{P}_X^{(0)}) \\ &= \nu\gamma_{\mathcal{F}}(\mathbb{P}_X^{(1)},\mathbb{P}_X^{(0)}) - \nu\gamma_{\mathcal{F}}(g\#\mathbb{P}_X^{(1)}, g\#\mathbb{P}_X^{(0)}) - \gamma_{\mathcal{F}}(g\#\mathbb{P}_X^{(0)},\mathbb{P}_X^{(0)}) \\ &\geq \nu(1-K)\gamma_{\mathcal{F}}(\mathbb{P}_X^{(1)},\mathbb{P}_X^{(0)}) - \gamma_{\mathcal{F}}(g\#\mathbb{P}_X^{(0)},\mathbb{P}_X^{(0)})\end{aligned}$$

for all $\nu \in (0,1]$. The first inequality follows from triangle inequality and the first equality follows from Lemma 1. The second inequality follows due to $g \in \mathcal{H}_{\mathcal{F}}^{(K)}$. Let $h^{(0)}$ be a function satisfying $T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h^{(0)}) \leq T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, g)$ for all $g \in \mathcal{H}_{\mathcal{F}}^{(K)}$. Then, by the assumption $T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h) < \epsilon\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)},\mathbb{P}_X^{(1)})$ for some $\epsilon < (1-K)/2$, we have $T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h^{(0)}) < \epsilon\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)},\mathbb{P}_X^{(1)})$. Thus, we have

$$\begin{aligned}&\nu(1-K)\gamma_{\mathcal{F}}(\mathbb{P}_X^{(1)},\mathbb{P}_X^{(0)}) - \gamma_{\mathcal{F}}(h^{(0)}\#\mathbb{P}_X^{(0)},\mathbb{P}_X^{(0)}) \\ &= \gamma_{\mathcal{F}}(h^{(0)}\#\mathbb{P}_X^{(0)},\mathbb{P}_X^{(0)}) + \nu(1-K)\gamma_{\mathcal{F}}(\mathbb{P}_X^{(1)},\mathbb{P}_X^{(0)}) - 2\gamma_{\mathcal{F}}(h^{(0)}\#\mathbb{P}_X^{(0)},\mathbb{P}_X^{(0)}) \\ &> \gamma_{\mathcal{F}}(h^{(0)}\#\mathbb{P}_X^{(0)},\mathbb{P}_X^{(0)}) + (\nu(1-K)-2\epsilon)\gamma_{\mathcal{F}}(\mathbb{P}_X^{(1)},\mathbb{P}_X^{(0)}).\end{aligned}$$

This concludes the second statement of Theorem 1. The first statement, the admissibility of anomaly detection algorithm, is followed by well choosing ν^* such that $1 \geq \nu^* > 2\epsilon/(1-K)$. □

Proof of Example 2. Sketch of Proof: For any $a \in \mathbb{R}^2$ such that $a^Ta = 1$, an autoencoder $h_{a,K}(x) := Kaa^Tx$ is in $\mathcal{H}_{\mathcal{F}}^{(K)}$, and satisfies the assumption of Theorem 1, $T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h_{a,K}) < (1-K)\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)},\mathbb{P}_X^{(1)})/4$ when $||\mu||_2 > 4\sqrt{1+(1-K)^2}/(1-K)$.

First, we prove that $h_{a,K} \in \mathcal{H}_{\mathcal{F}}^{(K)}$. For any two probability measures $\mathbb{P}$ and $\mathbb{Q}$, we denote the set of all couplings of $\mathbb{P}$ and $\mathbb{Q}$ by $\Pi(\mathbb{P},\mathbb{Q})$. For any x_1 and x_2 in $\mathbb{R}^2$, $||h_{a,K}(x_1)-h_{a,K}(x_2)||_2 = K|a^T(x_1-x_2)| \leq K\sqrt{(a^Ta)(x_1-x_2)^T(x_1-x_2)} = K||x_1-x_2||_2$ holds, so $h_{a,K}$ is K-Lipschitz continuous w.r.t. d. Since $\mathcal{F}$ is $\mathcal{F}_d$, $\gamma_{\mathcal{F}}$ is the 1-Wasserstein distance w.r.t. d, and by the Kantorovich-Rubinstein duality (Villani, 2008), $\gamma_{\mathcal{F}}(h\#\mathbb{P}, h\#\mathbb{Q}) \leq \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})}\int_{\mathcal{X}\times\mathcal{X}}||h(x_1)-h(x_2)||_2 d\pi(x_1,x_2)$ for all $h \in \mathcal{H}_{\mathcal{F}}^{(K)}$. Thus, $\inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})}\int_{\mathcal{X}\times\mathcal{X}}||h_{a,K}(x_1)-h_{a,K}(x_2)||_2 d\pi(x_1,x_2) \leq \inf_{\pi\in\Pi(\mathbb{P},\mathbb{Q})}\int_{\mathcal{X}\times\mathcal{X}}K||x_1-x_2||_2 d\pi(x_1,x_2) = K\gamma_{\mathcal{F}}(\mathbb{P},\mathbb{Q})$ implies that $h_{a,K}$ is in $\mathcal{H}_{\mathcal{F}}^{(K)}$.

Next, we provide an upper bound of $T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h_{a,K})$ and a lower bound of $\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)},\mathbb{P}_X^{(1)})$. Since $X^{(0)} \sim N_2(0_2, I_2)$ and $h_{a,K}(X^{(0)}) \sim N_2(0_2, K^2aa^T)$, we can apply the closed form of 2-Wasserstein distance between two Gaussian random variables (Givens et al., 1984). This gives the 2-Wasserstein distance of $\sqrt{1+(1-K)^2}$, which is an upper bound of the 1-Wasserstein distance between $\mathbb{P}_X^{(0)}$ and $h_{a,K}\#\mathbb{P}_X^{(0)}$, *i.e.*, $T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h_{a,K})$. For any two random variables X and Y, $\mathbb{E}||X-Y||_2 \geq ||\mathbb{E}[X-Y]||_2$ by Jensen's inequality, so $\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)},\mathbb{P}_X^{(1)})$ has a lower bound of $||\mu||_2$. Thus, by the above bounds, $||\mu||_2 > 4\sqrt{1+(1-K)^2}/(1-K)$ implies $T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h_{a,K}) < (1-K)\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)},\mathbb{P}_X^{(1)})/4$. □

Proof of Theorem 2. By triangle inequality of $\gamma_{\mathcal{F}}$, K-Lipschitz continuity w.r.t. $\gamma_{\mathcal{F}}$, and the assumption $T_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h^{(0)}) < \epsilon\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)},\mathbb{P}_X^{(1)})$, we have

$$\begin{aligned}\gamma_{\mathcal{F}}(\delta_{x'}, h^{(0)}\#\delta_{x'}) &\leq \gamma_{\mathcal{F}}(\delta_{x'},\mathbb{P}_X^{(0)}) + \gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h^{(0)}\#\mathbb{P}_X^{(0)}) + \gamma_{\mathcal{F}}(h^{(0)}\#\mathbb{P}_X^{(0)}, h^{(0)}\#\delta_{x'}) \\ &\leq (1+K)\gamma_{\mathcal{F}}(\delta_{x'},\mathbb{P}_X^{(0)}) + \gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h^{(0)}\#\mathbb{P}_X^{(0)}) \\ &< (1+K)\gamma_{\mathcal{F}}(\delta_{x'},\mathbb{P}_X^{(0)}) + \epsilon\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)},\mathbb{P}_X^{(1)}) \\ &< (1+K)\gamma_{\mathcal{F}}(\delta_{x'},\mathbb{P}_X^{(0)}) + (1-K)\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)},\mathbb{P}_X^{(1)})/2.\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}\gamma_{\mathcal{F}}(\delta_x, h^{(0)}\#\delta_x) &\geq \gamma_{\mathcal{F}}(\delta_x, \mathbb{P}_X^{(0)}) - \gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h^{(0)}\#\mathbb{P}_X^{(0)}) - \gamma_{\mathcal{F}}(h^{(0)}\#\mathbb{P}_X^{(0)}, h^{(0)}\#\delta_x) \\ &\geq (1-K)\gamma_{\mathcal{F}}(\delta_x, \mathbb{P}_X^{(0)}) - \gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, h^{(0)}\#\mathbb{P}_X^{(0)}) \\ &> (1-K)\gamma_{\mathcal{F}}(\delta_x, \mathbb{P}_X^{(0)}) - \epsilon\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)}) \\ &> (1-K)\gamma_{\mathcal{F}}(\delta_x, \mathbb{P}_X^{(0)}) - (1-K)\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)})/2.\end{aligned}$$

Since

$$\begin{aligned}&\gamma_{\mathcal{F}}(\delta_x, \mathbb{P}_X^{(0)}) > (1+K)/(1-K)\gamma_{\mathcal{F}}(\delta_{x'}, \mathbb{P}_X^{(0)}) + \gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)}) \\ \Longleftrightarrow &(1-K)\gamma_{\mathcal{F}}(\delta_x, \mathbb{P}_X^{(0)}) > (1+K)\gamma_{\mathcal{F}}(\delta_{x'}, \mathbb{P}_X^{(0)}) + (1-K)\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)}) \\ \Longleftrightarrow &(1-K)\gamma_{\mathcal{F}}(\delta_x, \mathbb{P}_X^{(0)}) - (1-K)\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)})/2 > (1+K)\gamma_{\mathcal{F}}(\delta_{x'}, \mathbb{P}_X^{(0)}) + (1-K)\gamma_{\mathcal{F}}(\mathbb{P}_X^{(0)}, \mathbb{P}_X^{(1)})/2,\end{aligned}$$

we have $T_{\mathcal{F}}(\delta_x, h^{(0)}) > T_{\mathcal{F}}(\delta_{x'}, h^{(0)})$. □

Proof of Proposition 4. By the definition of IPM,

$$\begin{aligned}&\gamma_{\mathcal{F}}(\nu\delta_x + (1-\nu)\mathbb{P}_X^{(0)}, \nu h\#\delta_x + (1-\nu)h\#\mathbb{P}_X^{(0)}) \\ &= \sup_{f\in\mathcal{F}} \left|\nu\big(f(x) - f(h(x))\big) + (1-\nu)\big(\mathbb{E}_{X\sim\mathbb{P}_X^{(0)}} f(X) - \mathbb{E}_{X\sim\mathbb{P}_X^{(0)}} f(h(X))\big)\right|.\end{aligned}$$

Then, the triangle inequality w.r.t. $|\cdot|$ implies

$$\begin{aligned}&\sup_{f\in\mathcal{F}} \left|\nu\big(f(x) - f(h(x))\big) + (1-\nu)\big(\mathbb{E}_{X\sim\mathbb{P}_X^{(0)}} f(X) - \mathbb{E}_{X\sim\mathbb{P}_X^{(0)}} f(h(X))\big)\right| \\ &\leq \nu \sup_{f\in\mathcal{F}} \big|f(x) - f(h(x))\big| + (1-\nu)\sup_{f\in\mathcal{F}} \left|\mathbb{E}_{X\sim\mathbb{P}_X^{(0)}} f(X) - \mathbb{E}_{X\sim\mathbb{P}_X^{(0)}} f(h(X))\right|,\end{aligned}$$

which concludes the proof. □

B Implementation details

B.1 Dataset description and preprocessing

KDD99: A large-scale network-traffic records dataset. We use KDD99 10 percent dataset that consists of 494,021 network-traffic records with 41 attributes. One-hot encoding is applied, and the last dimension of each encoded vector is dropped, yielding 115-dimensional input data. After that, the dataset is rescaled to $[-1, 1]$ by min-max scaling.

MNIST: An image dataset consists of 70,000 images of handwritten digits in 28×28 gray-scale, and each image is labeled among one of 10 classes of digits. The resolution is resized to 32×32. The dataset is rescaled to $[0, 1]$ by min-max scaling.

Fashion-MNIST: An image dataset consists of 70,000 images of fashion products in 28×28 gray-scale, and each image is labeled among one of 10 classes of product type such as dress and coat. The resolution is resized to 32×32, and the dataset is rescaled to $[0, 1]$ by min-max scaling.

CelebA: A large-scale image dataset consists of 202,599 celebrity face images, and each image has 40 binary attribute annotations about appearance such as wearing eyeglasses. We use randomly sampled 25,000 images of male celebrities. For each image, we first cropped the 140×140 pixels on the center part and then resized into 64×64. The dataset is rescaled to $[-1, 1]$ by min-max scaling.

B.2 Network configuration

In this subsection, we provide a detailed configuration of network architecture and hyperparameters. For all methods, the batch size is 50 for KDD99 and 100 for MNIST, Fashion-MNIST, and CelebA. The number of epochs is 200 for KDD99, 50 for MNIST and Fashion-MNIST, and 100 for CelebA. The early stopping with ten patience is applied.

Table 1: The architecture of the proposed method for KDD99.

Operation	Input unit	Output unit
Encoder		
Dense-ReLU	115	30
Dense	30	5
Decoder		
Dense-ReLU	5	30
Dense-Tanh	30	115
Optimizer	Adam(β_1=0.9,β_2=0.999)	
Latent dimension	5	
Learning rate	2×10^{-4}	

Proposed method: We use the kernel $k(x, y) = \sum_{c\in C} 2d_z c/(2d_z c + ||x - y||_2^2)$ where d_z is the dimension of the latent space and $C = \{0.2, 0.5, 1, 2, 5\}$. The $\mathbb{P}_Z$ is set to standard Gaussian distribution with dimension d_z. We set (λ, ϕ, K) to be $(0.0, 10.0, 0.7)$ for KDD99 and $(2.0, 2.0, 0.8)$ for other datasets. Table 1 and Table 2 presents architectures for KDD99 and other datasets, respectively.

Deep SVDD: For KDD99 and CelebA, the architecture of deep SVDD is the same as the encoder part of the proposed method. For MNIST and Fashion-MNIST, we utilize the architecture and hyperparameters suggested by the Ruff et al. (2018). For CelebA, we use parameters used in MNIST and Fashion-MNIST.

ALAD: For KDD99, the suggested architecture and hyperparameters on Zenati et al. (2018) are used. Table 3 shows the architecture for MNIST and Fashion-MNIST, and Table 4 shows the architecture for CelebA. For CelebA, we use parameters used in MNIST and Fashion-MNIST.

Table 2: The architecture of the proposed method for MNIST, Fashion-MNIST, and CelebA.

re height	Operation	Kernel	Strides	Filter size	Batch normalization
Encoder					
	Conv-ReLU	5×5	1×1	F	✓
	Conv-ReLU	5×5	2×2	2F	✓
	Conv-ReLU	5×5	2×2	4F	✓
	Conv-ReLU	5×5	2×2	8F	✓
	Dense				✗
Decoder					
	Dense				✗
	Transpose Conv-ReLU	5×5	2×2	8F	✓
	Transpose Conv-ReLU	5×5	2×2	4F	✓
	Transpose Conv-ReLU	5×5	2×2	2F	✓
	Transpose Conv*	5×5	1×1	F	

Optimizer	Adam (β_1=0.9,β_2=0.999)
Learning rate	2×10^{-4}
Latent dimension	MNIST/Fashion-MNIST: 8 CelebA: 64
Filter size	MNIST/Fashion-MNIST: F=16 CelebA: F=64
Transpose Conv*	MNIST/Fashion-MNIST: Transpose Conv-Sigmoid CelebA: Transpose Conv-Tanh

Table 3: The architecture of ALAD for MNIST and Fashion-MNIST.

	Operation	Kernel	Strides	Filter size	Batch normalization
Encoder					
	Conv-LReLU	5×5	2×2	64	✓
	Conv-LReLU	5×5	2×2	128	✓
	Conv-LReLU	5×5	2×2	256	✓
	Conv	4×4	1×1	8	✗
Generator					
	Transpose Conv-ReLU	8×8	2×2	256	✓
	Transpose Conv-ReLU	5×5	2×2	128	✓
	Transpose Conv-ReLU	5×5	2×2	64	✓
	Transpose Conv-Tanh	5×5	1×1	3	✓
Discriminator for (X, Z)					
only on X	Conv-LReLU	4×4	2×2	64	✗
	Conv-LReLU	4×4	2×2	128	✓
	Conv-LReLU	4×4	2×2	256	✓
only on Z	Conv-LReLU	1×1	1×1	256	✗
	Conv-LReLU	1×1	1×1	256	✗
concat outputs	Conv-LReLU	1×1	1×1	512	✗
	Conv-LReLU	1×1	1×1	1	✗
Discriminator for (X, X')					
concat X, X'	Conv-LReLU	5×5	2×2	32	✗
	Conv-LReLU	5×5	2×2	64	✗
	Dense			1	✗
Discriminator for (Z, Z')					
concat Z, Z'	Dense-LReLU			32	✗
	Dense-LReLU			16	✗
	Dense-LReLU			1	✗

Table 4: The architecture of ALAD for CelebA.

Operation		Kernel	Strides	Filter size	Batch normalization
Encoder					
	Conv-LReLU	5×5	2×2	64	✓
	Conv-LReLU	5×5	2×2	128	✓
	Conv-LReLU	5×5	2×2	256	✓
	Conv-LReLU	5×5	2×2	512	✓
	Conv	4×4	1×1	64	✗
Generator					
	Transpose Conv-ReLU	8×8	2×2	512	✗
	Transpose Conv-ReLU	5×5	2×2	256	✓
	Transpose Conv-ReLU	5×5	2×2	128	✓
	Transpose Conv-ReLU	5×5	2×2	64	✓
	Transpose Conv-Tanh	5×5	1×1	3	✓
Discriminator for (X, Z)					
only on X	Conv-LReLU	4×4	2×2	128	✗
	Conv-LReLU	4×4	2×2	256	✓
	Conv-LReLU	4×4	2×2	512	✓
only on Z	Conv-LReLU	1×1	1×1	512	✗
	Conv-LReLU	1×1	1×1	512	✗
concat outputs	Conv-LReLU	1×1	1×1	1024	✗
	Conv-LReLU	1×1	1×1	1	✗
Discriminator for (X, X')					
concat X, X'	Conv-LReLU	5×5	2×2	64	✗
	Conv-LReLU	5×5	2×2	128	✗
	Conv-LReLU	5×5	2×2	256	✗
	Dense			1	✗
Discriminator for (Z, Z')					
concat Z, Z'	Dense-LReLU			64	✗
	Dense-LReLU			32	✗
	Dense-LReLU			1	✗

C Detailed results of ablation study

Table 5: Average AUCs on KDD99 of the proposed method for various λ, ϕ, and K are provided in % with standard deviation. The number of replication is 10. For each level of λ, the row where ϕ is 0 presents the baseline performance when Lipschitz continuity is not enforced.

ϕ	K								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\lambda = 0$									
0	98.6±0.6	98.6±0.6	98.6±0.6	98.6±0.6	98.6±0.6	98.6±0.6	98.6±0.6	98.6±0.6	98.6±0.6
5	98.9±0.2	99.0±0.4	99.0±0.2	99.1±0.3	99.1±0.2	99.2±0.1	99.3±0.1	99.3±0.1	98.1±1.1
10	99.1±0.2	99.1±0.3	99.0±0.3	99.1±0.3	99.2±0.3	99.2±0.1	**99.4±0.1**	96.9±1.3	95.3±2.4
20	98.7±0.2	98.8±0.2	99.1±0.2	99.2±0.9	99.2±0.2	96.6±0.6	89.4±1.6	91.9±0.3	96.8±1.8
$\lambda = 5$									
0	98.2±1.0	98.2±1.0	98.2±1.0	98.2±1.0	98.2±1.0	98.2±1.0	98.2±1.0	98.2±1.0	98.2±1.0
5	98.7±0.6	98.2±0.6	98.8±0.4	98.5±0.6	**99.1±0.4**	98.3±1.1	98.8±0.3	98.6±0.4	97.6±1.5
10	98.3±0.3	98.7±0.4	98.7±0.5	98.6±0.3	98.7±0.4	98.9±0.5	98.7±0.4	96.4±2.1	97.3±2.1
20	98.6±0.4	98.9±0.4	98.7±0.3	98.8±0.3	98.6±0.4	95.9±2.2	92.9±3.9	94.9±3.2	97.0±5.4
$\lambda = 10$									
0	97.2±1.1	97.2±1.1	97.2±1.1	97.2±1.1	97.2±1.1	97.2±1.1	97.2±1.1	97.2±1.1	97.2±1.1
5	98.6±0.6	98.9±0.7	98.4±1.3	98.7±0.6	98.2±0.5	98.5±0.4	98.8±0.5	98.3±1.3	97.4±1.9
10	98.4±0.4	98.7±0.5	98.9±0.5	98.8±0.4	98.6±0.4	98.7±0.3	98.7±0.4	97.2±1.3	96.5±2.0
20	98.6±0.2	98.7±0.3	98.8±0.2	**99.0±0.2**	98.7±0.4	96.8±1.1	93.3±4.2	94.1±3.3	97.4±1.6

Table 5 and 6 shows AUCs and AUPRCs, respectively, of ablation study to evaluate the effect of Lipschitz continuity imposed on autoencoders. The penalty term to enforce Lipschitz continuity has hyperparameter K with a coefficient ϕ, and the level of enforcement increases as ϕ increases or K decreases. The ablation study is conducted with the uncontaminated dataset. Compared with the baseline model, a moderate level of Lipschitz continuity significantly enhances the performance for every λ. For $\lambda = 0$, the mean AUC and AUPRC is 98.6 with std of 0.6 and 93.0 with std of 2.6 from the baseline model and increases to 99.4 with std of 0.1 and 96.4 with std of 0.3, respectively. For $\lambda = 5$, the mean AUC and AUPRC is 98.2 with std of 1.0 and 91.3 with std of 3.2 from the baseline model and increases to 99.1 with std of 0.4 and 94.0 with std of 2.2, respectively. For $\lambda = 10$, the mean AUC and AUPRC is 97.2 with std of 1.1 and 88.0 with std of 3.1 from the baseline model and increases to 99.0 with std of 0.2 and 92.8 with std of 2.4, respectively.

Table 6: Average AUPRCs on KDD99 of the proposed method for various λ, ϕ, and K are provided in % with standard deviation. The number of replication is 10. For each level of λ, the row where ϕ is 0 presents the baseline performance when Lipschitz continuity is not enforced.

ϕ					K				
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\lambda = 0$									
0	93.0±2.6	93.0±2.6	93.0±2.6	93.0±2.6	93.0±2.6	93.0±2.6	93.0±2.6	93.0±2.6	93.0±2.6
5	94.5±0.9	94.6±1.8	94.9±1.3	95.3±1.8	95.3±0.9	95.4±0.8	96.2±0.8	95.8±0.8	92.8±2.3
10	95.1±0.9	94.8±1.9	94.8±1.8	94.9±1.4	95.4±1.4	95.6±0.6	**96.4±0.3**	89.6±2.4	88.0±4.4
20	92.2±1.6	93.4±1.3	94.9±1.4	95.6±0.9	95.4±1.2	86.5±1.9	73.8±3.8	81.0±0.8	89.9±3.1
$\lambda = 5$									
0	91.3±3.2	91.3±3.2	91.3±3.2	91.3±3.2	91.3±3.2	91.3±3.2	91.3±3.2	91.3±3.2	91.3±3.2
5	92.6±2.9	90.6±2.2	92.8±2.1	91.2±3.0	**94.0±2.2**	91.1±3.8	92.7±1.2	91.3±2.2	89.1±3.7
10	90.4±1.9	92.2±2.4	92.1±2.7	91.8±1.7	92.2±1.8	93.2±2.5	92.2±2.2	86.5±4.8	89.5±4.7
20	91.8±1.7	92.6±2.0	92.2±1.8	92.9±2.0	91.8±1.9	84.1±5.1	80.4±8.3	84.7±7.0	89.0±5.4
$\lambda = 10$									
0	88.0±3.1	88.0±3.1	88.0±3.1	88.0±3.1	88.0±3.1	88.0±3.1	88.0±3.1	88.0±3.1	88.0±3.1
5	91.9±2.5	93.2±3.5	91.9±3.3	92.1±2.8	89.8±2.6	91.1±2.6	**92.8±2.4**	91.7±3.2	90.0±5.4
10	90.5±2.0	91.9±2.6	**92.8±2.4**	92.1±2.0	91.4±2.0	92.3±1.6	92.4±1.9	87.7±3.2	87.7±4.9
20	91.6±1.4	91.5±2.0	92.2±1.4	94.0±1.3	92.2±2.3	85.2±3.9	80.4±9.0	82.8±7.4	89.2±3.5

D Experiment results on MNIST and Fashion-MNIST

Table 7: Average AUCs of deep SVDD, ALAD, and the proposed method are provided in % with standard deviation. The number of replication is 10.

Dataset	Normal class	Uncontaminated Training Set			Contaminated Training Set		
		deep SVDD	ALAD	Proposed	deep SVDD	ALAD	Proposed
MNIST (32x32)	Digit 0	97.90±0.8	98.83±0.7	**98.86±0.2**	94.40±2.0	95.50±2.0	**96.64±0.4**
	Digit 1	99.47±0.2	99.22±0.4	**99.68±0.1**	98.76±0.3	99.12±0.5	**99.52±0.1**
	Digit 2	87.23±2.2	79.79±10.1	**90.45±1.6**	82.65±3.3	73.25±13.3	**84.30±2.4**
	Digit 3	89.91±1.7	85.32±3.2	**93.44±0.7**	86.72±2.9	82.09±3.8	**88.25±2.3**
	Digit 4	93.06±1.4	91.4±2.3	**95.70±0.6**	90.04±1.5	89.37±2.6	**92.83±0.8**
	Digit 5	88.31±2.1	83.18±11.1	**94.54±1.3**	83.13±1.9	83.61±2.6	**90.56±1.8**
	Digit 6	98.17±0.5	93.81±11.8	**98.43±0.5**	95.07±1.1	**96.55±2.5**	96.28±0.8
	Digit 7	93.98±1.5	94.38±1.3	**96.39±0.8**	90.77±1.8	92.53±2.3	**94.25±0.5**
	Digit 8	**89.94±1.9**	86.38±5.1	89.81±1.4	87.33±2.6	**88.10±2.7**	83.75±2.4
	Digit 9	**96.05±0.6**	95.49±1.4	95.67±0.6	93.94±0.6	**94.79±1.2**	94.28±1.1
Fashion- MNIST (32x32)	T-shirt/top	90.66±0.9	89.45±9.0	**92.73±0.5**	86.94±2.0	87.33±1.2	**88.04±1.2**
	Trouser	**98.65±0.1**	98.47±0.2	98.52±0.2	**97.57±0.3**	90.35±21.8	97.38±0.3
	Pullover	86.22±2.9	89.62±0.9	**89.72±0.6**	83.90±2.4	**86.07±1.7**	85.95±1.0
	Dress	92.62±1.3	58.01±23.6	**94.47±0.3**	90.17±1.3	67.11±24.7	**91.18±1.3**
	Coat	89.31±2.6	89.62±1.5	**91.41±0.3**	87.34±1.7	87.96±0.8	**88.15±0.7**
	Sandal	90.03±1.1	23.07±9.1	**91.54±0.3**	81.13±1.5	24.41±9.9	**83.44±2.0**
	Shirt	80.62±1.9	84.10±0.6	**84.80±0.6**	79.14±1.2	**80.44±1.4**	79.94±1.4
	Sneaker	98.48±0.1	27.56±24.7	**98.50±0.1**	97.13±0.5	50.95±37.6	**97.43±0.4**
	Bag	**92.06±2.9**	85.05±3.0	90.59±1.3	83.72±2.4	**83.75±2.4**	77.73±2.1
	Ankle boot	**98.28±0.3**	92.39±2.3	98.12±0.2	**94.52±1.2**	83.58±13.6	92.87±1.5

Table 7 and Table 8 shows AUCs and AUPRCs of various methods on MNIST and Fashion-MNIST dataset, respectively. All the mean and standard deviation values are based on 10 runs. In each table, we mark the highest value in bold for both uncontaminated and contaminated training set cases.

Table 8: Average AUPRCs of deep SVDD, ALAD, and the proposed method are provided in % with standard deviation. The number of replication is 10.

Dataset	Normal class	Uncontaminated Training Set			Contaminated Training Set		
		deep SVDD	ALAD	Proposed	deep SVDD	ALAD	Proposed
MNIST (32x32)	Digit 0	99.73±0.1	**99.84±0.1**	**99.84±0.0**	99.17±0.3	99.42±0.3	**99.50±0.1**
	Digit 1	99.91±0.0	99.87±0.1	**99.95±0.0**	99.75±0.1	99.85±0.1	**99.92±0.0**
	Digit 2	98.22±0.3	96.93±2.1	**98.66±0.3**	97.47±0.6	95.73±3.0	**97.77±0.3**
	Digit 3	98.59±0.3	97.92±0.5	**99.08±0.1**	98.06±0.4	97.41±0.5	**98.31±0.3**
	Digit 4	99.15±0.2	98.95±0.3	**99.45±0.1**	98.62±0.3	98.68±0.4	**99.07±0.1**
	Digit 5	98.48±0.4	97.70±1.9	**99.30±0.2**	97.71±0.2	97.84±0.5	**98.80±0.2**
	Digit 6	99.76±0.1	98.97±2.2	**99.78±0.1**	99.29±0.2	**99.51±0.4**	99.47±0.1
	Digit 7	99.18±0.2	99.26±0.2	**99.52±0.1**	98.60±0.3	99.01±0.3	**99.20±0.1**
	Digit 8	**98.66±0.3**	97.96±0.8	98.64±0.2	**98.26±0.4**	98.11±0.6	97.77±0.4
	Digit 9	**99.47±0.1**	99.37±0.2	99.37±0.1	99.13±0.1	**99.25±0.2**	99.15±0.2
Fashion-MNIST (32x32)	T-shirt/top	98.75±0.1	98.28±1.9	**99.01±0.1**	**98.04±0.3**	97.97±0.2	98.01±0.2
	Trouser	**99.83±0.0**	99.79±0.0	99.77±0.0	**99.62±0.0**	97.8±5.5	99.58±0.1
	Pullover	98.32±0.4	**98.76±0.1**	98.74±0.1	97.88±0.3	**98.14±0.3**	97.86±0.2
	Dress	99.05±0.2	91.42±5.2	**99.29±0.0**	98.55±0.4	93.44±5.5	**98.76±0.2**
	Coat	98.69±0.3	98.70±0.2	**98.95±0.0**	98.29±0.2	**98.31±0.2**	98.26±0.1
	Sandal	98.88±0.1	82.51±4.4	**99.01±0.0**	**97.38±0.2**	83.05±4.6	97.35±0.4
	Shirt	97.27±0.3	97.83±0.1	**97.91±0.1**	96.86±0.3	**97.06±0.3**	96.74±0.3
	Sneaker	**99.84±0.0**	84.84±5.9	99.83±0.0	99.62±0.1	90.14±8.0	**99.65±0.1**
	Bag	**98.92±0.4**	97.75±0.4	98.71±0.1	**97.55±0.3**	97.54±0.4	96.55±0.4
	Ankle boot	**99.80±0.0**	99.08±0.3	**99.78±0.0**	99.23±0.1	97.48±2.6	98.9±0.2

References

Givens, C. R., Shortt, R. M., et al. (1984). A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240.

Ruff, L., Görnitz, N., Deecke, L., Siddiqui, S. A., Vandermeulen, R., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In *International Conference on Machine Learning*, pages 4390–4399.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Zenati, H., Romain, M., Foo, C.-S., Lecouat, B., and Chandrasekhar, V. (2018). Adversarially learned anomaly detection. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 727–736. IEEE.