Supplementary Material

1 Proofs

Proof of Equivalence in Definition 2.1 in the main paper. While the proof of the equivalence is well-known, we reproduce here for completeness. As a reminder, the KL divergence is defined as:

$$\operatorname{KL}(P,Q) = \mathbb{E}_P[\log \frac{P(z)}{Q(z)}] = \mathbb{E}_P[-\log Q(z)] - \mathbb{E}_P[-\log P(z)] = \operatorname{H}_c(P,Q) - \operatorname{H}(P), \quad (1)$$

where $H_c(\cdot,\cdot)$ denotes the cross entropy and $H(\cdot)$ denotes entropy. Given this, we can now easily derive the equivalence:

$$GJSD_{\boldsymbol{w}}(P_{Z_1},\cdots,P_{Z_k}) = \sum_j w_j \operatorname{KL}(P_{Z_j},P_{Z_{\operatorname{mix}}})$$
(2)

$$= \sum_{j} w_j (\mathrm{H}_{\mathrm{c}}(P_{Z_j}, P_{Z_{\mathrm{mix}}}) - \mathrm{H}(P_{Z_j}))$$
(3)

$$= \sum_{j} w_{j} \operatorname{H}_{c}(P_{Z_{j}}, P_{Z_{\text{mix}}}) - \sum_{j} w_{j} \operatorname{H}(P_{Z_{j}})$$
(4)

$$= \sum_{j} w_{j} \mathbb{E}_{P_{Z_{j}}} [-\log P_{Z_{\text{mix}}}] - \sum_{j} w_{j} \operatorname{H}(P_{Z_{j}})$$
(5)

$$= \sum_{j} w_j \int_{\mathcal{Z}} -P_{Z_j}(z) \log P_{Z_{\text{mix}}}(z) dz - \sum_{j} w_j \operatorname{H}(P_{Z_j})$$
(6)

$$= \int_{\mathcal{Z}} -\sum_{j} w_{j} P_{Z_{j}}(z) \log P_{Z_{\text{mix}}}(z) dz - \sum_{j} w_{j} \operatorname{H}(P_{Z_{j}})$$
(7)

$$= \int_{\mathcal{Z}} -P_{Z_{\text{mix}}}(z) \log P_{Z_{\text{mix}}}(z) dz - \sum_{j} w_{j} \operatorname{H}(P_{Z_{j}})$$
(8)

$$= \mathrm{H}(P_{Z_{\mathrm{mix}}}) - \sum_{j} w_{j} \mathrm{H}(P_{Z_{j}}) \,. \tag{9}$$

Proof of Lemma 2.3 in the main paper. First, we note the following fact from the standard change of variables formula:

$$P_X(\boldsymbol{x}) = P_Z(T(\boldsymbol{x}))|J_T(\boldsymbol{x})|$$

$$\Rightarrow P_X(\boldsymbol{x})|J_T(\boldsymbol{x})|^{-1} = P_Z(T(\boldsymbol{x})).$$
(10)

We can now derive our result using the change of variables for expectations (i.e., LOTUS) and the probability change of variables from above:

$$\begin{split} \mathrm{H}(P_Z) &= \mathbb{E}_{P_Z}[-\log P_Z(\boldsymbol{z})] = \mathbb{E}_{P_X}[-\log P_Z(T(\boldsymbol{x}))] \\ &= \mathbb{E}_{P_X}[-\log P_X(\boldsymbol{x})|J_T(\boldsymbol{x})|^{-1}] \\ &= \mathbb{E}_{P_X}[-\log P_X(\boldsymbol{x})] + \mathbb{E}_{P_X}[-\log |J_T(\boldsymbol{x})|^{-1}] \\ &= \mathrm{H}(P_X) + \mathbb{E}_{P_X}[\log |J_T(\boldsymbol{x})|] \,. \end{split}$$

Proof of Theorem 2.5 in the main paper. Given any fixed Q, minimizing \mathcal{L}_{AUB} decouples into minimizing separate normalizing flow losses where Q is the base distribution. For each normalizing flow, there exists an invertible T_j such that $T_j(X_j) \sim Q$, and this achieves the minimum value of \mathcal{L}_{AUB} .

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

More formally,

$$\min_{T_1,\cdots,T_k} \mathcal{L}_{\text{AUB}}(T_1,\cdots,T_k) \tag{11}$$

$$= \min_{T_1, \cdots, T_k} \sum_j w_j \mathbb{E}_{P_{X_j}} \left[-\log \left| J_{T_j}(\boldsymbol{x}) \right| Q(T_j(\boldsymbol{x})) \right]$$
(12)

$$= \sum_{j} w_{j} \min_{T_{j}} \mathbb{E}_{P_{X_{j}}} \left[-\log |J_{T_{j}}(\boldsymbol{x})| Q(T_{j}(\boldsymbol{x})) \right] + \mathrm{H}(P_{X_{j}}) - \mathrm{H}(P_{X_{j}})$$
(13)

$$= \sum_{j} w_{j} \min_{T_{j}} \mathbb{E}_{P_{X_{j}}} \left[-\log |J_{T_{j}}(\boldsymbol{x})| Q(T_{j}(\boldsymbol{x})) \right] + \mathcal{H}(P_{X_{j}}) - \mathbb{E}_{P_{X_{j}}} \left[-\log P_{X_{j}}(\boldsymbol{x}) \right] \right)$$
(14)

$$= \sum_{j} w_{j} \operatorname{H}(P_{X_{j}}) + \sum_{j} w_{j} \min_{T_{j}} \mathbb{E}_{P_{X_{j}}} \left[\log \frac{P_{X_{j}}(\boldsymbol{x}) |J_{T_{j}}(\boldsymbol{x})|^{-1}}{Q(T_{j}(\boldsymbol{x}))} \right]$$
(15)

$$= \sum_{j} w_{j} \operatorname{H}(P_{X_{j}}) + \sum_{j} w_{j} \min_{T_{j}} \mathbb{E}_{P_{X_{j}}} \left[\log \frac{P_{T_{j}(X_{j})}(T_{j}(\boldsymbol{x}))}{Q(T_{j}(\boldsymbol{x}))} \right]$$
(16)

$$= \sum_{j} w_{j} \operatorname{H}(P_{X_{j}}) + \sum_{j} w_{j} \min_{T_{j}} \mathbb{E}_{P_{T_{j}}(X_{j})} \left[\log \frac{P_{T_{j}}(X_{j})(z)}{Q(z)} \right]$$
(17)

$$= \sum_{j} w_j \operatorname{H}(P_{X_j}) + \sum_{j} w_j \min_{T_j} \operatorname{KL}(P_{T_j(X_j)}, Q).$$
(18)

Given that $\operatorname{KL}(P,Q) \ge 0$ and equal to 0 if and only if P = Q, the global minimum is achieved only if $P_{T_j(X_j)} = Q, \forall j$ and there exist such invertible functions (e.g., the optimal Monge map between P_{X_j} and Q for squared Euclidean transportation cost [1]). Additionally, the optimal value is $\sum_j w_j \operatorname{H}(P_{X_j})$, which is constant with respect to the T_j transformations. \Box

2 Additional Experiment Details

2.1 Toy Dataset Experiment

LRMF vs. Ours Experiment

- T for LRMF setup: T_1 : 8 channel-wise mask for Real-NVP model with s and t derived from 64 hidden channels of fully connected networks. T_2 : Identity function.
- T for RAUB setup: T_1 and T_1 : 8 channel-wise mask for Real-NVP model with s and t derived from 64 hidden channels of fully connected networks.
- Q for both: A single Gaussian distribution with trainable mean and trainable variances.

Alignflow vs. Ours Experiment

- T for both: 2 channel-wise mask for RealNVP model with s and t derived from 8 hidden channels of fully connected networks.
- Q for Alignflow setup: A single fixed normal distribution.
- Q for RAUB setup: A learnable mixture of Gaussian with 3 components.

2.2 Tabular Dataset Experiment

Our invertible transformation function T adapts general purpose RealNVP (5) and RealNVP (10) layers with detailed parameters provided in Table 1. Successive coupling layers are concatenated by alternating the format between keeping odd-parity index of the data samples and transforms the even-parity index of the data and vice versa. Note that our density model Q for AUB experiments shares the same architecture of RealNVP (5) throughout the two-domain tabular dataset experiment and RealNVP (10) for multi-domain experiment. For AlignFlow hybrid and adversarial only models, in order to adapt their model for a tabular dataset, we change the invertible model T_src and T_tgt to RealNVP(5) and change the discriminator to a fully connected network which has a hidden dimension of 256.

3 Multi-domain translation

To illustrate that our method can be easily scaled to more domain distributions even for high dimensional data, we present qualitative examples of translating between every digit and every other

Table 1: RealNVP layers used in tabular experiment section contains $\langle n_layers \rangle$ coupling layers. Each coupling layer has two fully connected networks modeling the scaling and shifting function with $\langle hidden_dim \rangle$ hidden dimensions and $\langle n_laden \rangle$ hidden layers.

	n_layers	hidden_dim	n_hidden	Total number of parameters
RealNVP(5)	5	100	1	1,462,400
RealNVP(10)	10	100	2	4,540,800

Table 2: FID and AUB score for domain alignment task in 10 domains. FID score is calculated by average across all paired translations and AUB score is shown in nats.

		FID	AUB				
	AlignFlow (MLE)	49.82	-4661.05				
	Ours	43.25	-4715.02				
01233333333333 1111111111111111111111111	45678999999999999999999999999999999999999		22222222222222222222222222222222222222	44444444444444444444444444444444444444	マフマフフフフフフマ	() らう ら ら う ち ち ち ち ち ち ち う	979999999999999
Align	FIOW (IVILE)			Ours			

Figure 1: Qualitative comparison on translation results across 10 classes with AlignFlow and ours.

digit for MNIST in Fig. 1 with quantitative performances in Table 2. Note that we omit LRMF in this experiment because the multi-domain situation is hard to deal with for LRMF setup due to LRMF's two distribution setup and asymptric model structure.

As shown in Table 2, our approach shows better performance in terms of FID and AUB than AlignFlow since our learnable density estimator can model more complex distribution than fixed simple density model in AlignFlow. In other words, the shared space of AlignFlow is limited because of the fixed simple density model.

The superiority of our method compared to AlignFlow in multiple-domain translation can also be verified through qualitative comparisons in Fig. 1. The leftmost column is an input image and the second and third macro columns are the results from AlignFlow and ours. By forwarding a given k-th latent $T_k(x_k)$ into 10 inverse transformation functions, respectively. It is easy to observe that ours has more clear results in most cases than baseline. Moreover, our model shows better performance in maintaining the original identity (e.g., width and type of a stroke) than the baseline, as seen in fifth, eighth and ninth rows. This is because we jointly train our transformation functions. This benefit of our approach may be crucial for other datasets such as human faces [2] where maintaining the original identity is important.

4 Implementation on Latent Space

Our model is also capable of generating samples in each domain. One needs to sample from the density model Q to have a latent image z first, and then forward to the inverse function T_i^{-1} to get

the image sampled in j^{th} domain. Examples of generated images for each domain in MNIST data are shown in Fig. 2. The quality of the generative result also reflects the tightness of the bound between the latent space and the latent density model as illustrated in Eqn. 4.



Figure 2: This figure shows the generated images of our model for all domains.

In order to visualize the benefit of our shared latent space, we further perform interpolation in the latent space. We first randomly select two distinct real images in one domain (in this case two '0's), and do a linear interpolation of the selected two images in the latent space. Then we translate all the interpolated images (including the two selected images) to all of the remaining domains to generate 'translated-interpolated' images, i.e., the corresponding interpolations in each of the remaining domains. As shown in Fig. 3, all 'translated-interpolated' results can preserve the trend of the stroke width of the digits from the original interpolated domain. These results suggest that our shared space aligns the domains so that some latent space directions have similar semantic meaning for all domains.



Figure 3: This figure shows the translation results of interpolated images. In the first row, the two images selected by red rectangles are real images from the dataset; and all eight images in between are generated by linear interpolation in the latent space. Starting from the second row, each row contains translated images which are transformed from the same latent vector in the same column and the first row.

5 Qualitative results of Domain Adaptation experiments on USPS-MNIST dataset

Here are the translated results that we used in Table 4. The results consistently show that ours show the best performance. Please note that these samples are randomly chosen.



Figure 4: Digit 0 results. From left, USPS data, LRMF, AF (Adv. only), AF (MLE), AF (1e0), AUB (ours)

(1))))) 7, 7,5342	333333333	11126111	4 8 8 8 8 8 8 8 4	1/////////
1141111	333333333	22666671	1161114	111111
11111111 22/22/2017	333333333	88111111	4 8 8 9 8 4 8 9	11111111
1111111 15 16 23 4	333333333	68119810		112/11//
1111111 200 200 200	333333333	28111111	191144444	11111111
11111111	333333333	71486182		11/1/11
))ししししし エアニンシアクト	333333333	51011569	11211111	111111
11111111 248 2725	333333333	01609118		1111111111

Figure 5: Digit 1 results. From left, USPS data, LRMF, AF (Adv. only), AF (MLE), AF (1e0), AUB (ours)

22222222	1031 NY 4	333333333	20935770	12222222	22422222
22-227-22	S. C. C. Star	33333333	00012097	27027022	57227022
22222322	C . () . () . ()	33333333	39297001	22821822	22222322
2222222	いいいよんちょう	333333333	70781304	22 = 22722	28-72222
2222222	State -	333333333	90570002	12795521	21222228
211222222	ti Xala	333333333	90799275	11122722	=1722200
222-2222	1.	333333335	47303370	22222222	22242252
22222222	51451 1.7	333333333	77638201	12122132	22224222

Figure 6: Digit 2 results. From left, USPS data, LRMF, AF (Adv. only), AF (MLE), AF (1e0), AUB (ours)

33333333	2. A A	333333333	04023935	3333353	53323553
339333333		33333333	52928233	53333333	52933333
333339 33		333333333	P.52 E1228	33323933	33339933
33 333333 33	1. 1. 4. 2. 6. 5. 5.	33333333	03383189	3 3 3 3 3 3 3 3 5	33533333
3333333333	45.87 F 257	333333333	26565392	33335335	33333333
233333333	1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	333333333	30338849	27333935	27355583
<u>33333333</u> 333	N. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.	33333333	78228433	58333333	53333535
73733393	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	333333333	83363423	433333383	5335333

Figure 7: Digit 3 results. From left, USPS data, LRMF, AF (Adv. only), AF (MLE), AF (1e0), AUB (ours)

44444444	シンマン クリーム	3333333333	40149094	+ 4 9 4 4 4 4 4 4	44744444
4444144444	せいいちいかれ	333333333	79192104	44444474	44497474
4-14444444	1.12. 1. A 7 W	333333333	97444411	44444444	9-149-44444
44444444	こくちょうちょういい	88888888	79444990	44449444	44449441
444444444	1 - 2 - 2 - 2 - 2 - 2	333333333	40994440	4444449999	44497947
44444444	2 2 12 - 14 is	333333333	71798569	448494444	44841499
41144444	シントー インベン	333333333	46191894	49144444	97444444
14444444	N 19 11 X 101- 1	333333333	30499140	3+444449	17499491

Figure 8: Digit 4 results. From left, USPS data, LRMF, AF (Adv. only), AF (MLE), AF (1e0), AUB (ours)

3353 33 88	DOM SCHERE	333333333	10356596	83568556	5555555b
3 53555 55	522 2001	333333333	33633433	35559555	15583556
353355555	31377555	333333333	33538388	199395555	55355555
853333 59	ALAS 555	88888888	43553332	83555359	85553351
55555555		333333333	339999996	1: 955565	15155565
353055 63	26229737	333333333	57301180	5555555	55875576
383585555	48800 H C K	333333333	53933509	55556575	85556732
5 53553553	E1050.04.5	2 2 2 2 2 2 2 2 2 2	04568535	55555555	55655355

Figure 9: Digit 5 results. From left, USPS data, LRMF, AF (Adv. only), AF (MLE), AF (1e0), AUB (ours)

66666566	1915 19 Sec.	11111111	45471949	666566668	66666968
66666666	271.5. 240	3333,333	976-5449	60606666	66666666
66666666	· 21 1 1. 51	333333323	40846969	666666666	66665666
66666666	-1:511-37-2	88888888	99974440	666666668	66666866
66666666	1. Constant of the second	333323333	60445144	66466666	06066666
66666666	Vis Laks	333333333	51959594	66666666	666888666
66466666	- V T 1. 4. 7. 7 2	3333333333	64469446	66455666	66456666
66666666	- H , R H A T	いいちいいちょう	9-46-599	66666666	66666666

Figure 10: Digit 6 results. From left, USPS data, LRMF, AF (Adv. only), AF (MLE), AF (1e0), AUB (ours)



Figure 11: Digit 7 results. From left, USPS data, LRMF, AF (Adv. only), AF (MLE), AF (1e0), AUB (ours)



Figure 12: Digit 8 results. From left, USPS data, LRMF, AF (Adv. only), AF (MLE), AF (1e0), AUB (ours)



Figure 13: Digit 9 results. From left, USPS data, LRMF, AF (Adv. only), AF (MLE), AF (1e0), AUB (ours)

References

- [1] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends*® *in Machine Learning*, 11(5-6):355–607, 2019.
- [2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.