Supplementary Materials for "LTF: A Label Transformation Framework for Correcting Label Shift"

Anonymous Authors[1]

A. Supplementary Material

A.1. Proof of Proposition 1

Proof. Let h^c be complementary function of h, such that $X = H(h(X), h^c(X))$. Let $X_1 = h(X)$ and $X_2 = h^c(X)$, then we have $p_{X,Y}(x, y) = p_{X_1,X_2,Y}(x_1, x_2, y)$. Using the decomposition $p_{X_1,X_2,Y}(x_1, x_2, y) = p_{X_1}(x_1)p_{X_2,Y|X_1}(x_2, y|x_1)$, we have

$$I(Y, X) = I(Y, X_1) + E_{X_1}[I(Y|X_1, X_2|X_1)], \quad (1)$$

where $I(\cdot, \cdot) \geq 0$ is the mutual information. By maximizing $I(Y, X_1)$, the best solution we can achieve is $I(Y, X_1) = I(Y, X)$, which implies $I(Y|X_1, X_2|X_1) = 0$. This means the conditional independence of Y and X_2 given X_1, i.e., $Y \perp\!\!\!\perp X_2|X_1$, which is equivalent to $Y \perp\!\!\!\perp X|h(X)$. Then it suffices to show that maximizing mutual information $I(Y, X_1)$ is equivalent to minimizing the cross-entropy loss or mean squared loss under some parametric assumptions.

We first expand the mutual information $I(Y, X_1)$ as

$$\begin{aligned} I(Y, X_1) &= H(Y) - H(Y|X_1) \\ &= H(Y) + \int p(y, x_1) \log p(y|x_1) dy dx_1. \end{aligned} \quad (2)$$

For regression problems, we use $q(y|x_1) = N(w^T x_1|\sigma^2)$ to approximate $p(y|x_1)$ and write (2) as

$$I(Y, X_1) \approx H(Y) - \frac{1}{2\sigma^2} \int p(y, x_1)(y - w^T x_1)^2 dy dx_1. \quad (3)$$

It is straightforward to see that maximizing $I(Y, X_1)$ is equivalent minimizing the mean squared loss. For classification, we use $q(y = k|x_1) = \frac{\exp w_k^T x_1}{\sum_{k'=1}^K \exp w_{k'}^T x_1}$ to approximate $p(y|x_1)$ and rewrite (2) as

$$I(Y, X_1) \approx H(Y) + \int \sum_{k=1}^K p(y = k, x_1) \frac{\exp w_k^T x_1}{\sum_{k'=1}^K \exp w_{k'}^T x_1} dx_1. \quad (4)$$

Therefore, maximizing $I(Y, X_1)$ is equivalent minimizing the cross-entropy loss for classification. □

A.2. Results of Fashion-MNIST and MNIST

A.2.1. Results of CIFAR-10

The details of experimental settings of CIFAR-10 are given at the table 1, and the results of CIFAR-10 are given at the main paper.

Classifier Details	
Architecture	Resnet-18
Batch Size	128
Training epochs	20
Optimizer	SGD
Learning Rate	1e-2
L2 Penalty Parameter	5e-4
Label Transformation Details	
Architecture	One-Layer Network
Label Influence Recovery Details	
Generator Architecture	BigGAN
Training Method	BigGAN (Brock et al., 2018)
Distribution Matching	
Optimizer	Adam
Learning Rate	8e-5
Training epochs	1000

Table 1. The experimental details on CIFAR-10 dataset.

A.2.2. Results of Fashion-MNIST

The details of experimental settings of Fashion-MNIST are given at the table 2. The MSE error of the estimated label weights P_Y^T/P_Y^S, accuracy and F1 score of FASHION-MNIST are shown as Figure 1, 2, 3.

A.2.3. Results of MNIST

The details of experimental settings of MNIST are given at the table 3. The MSE error of the estimated label weights P_Y^T/P_Y^S, accuracy and F1 score of MNIST are shown as Figure 4, 5, 6.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Classifier Details	
Architecture	Discriminator of DCGAN
Batch Size	128
Training epochs	20
Optimizer	SGD
Learning Rate	1e-2
L2 Penalty Parameter	5e-4
Label Transformation Details	
Architecture	One-Layer Network
Label Influence Recovery Details	
Generator Architecture	Generator of DCGAN
Training Method	TAC-GAN (Gong et al., 2019)
Distribution Matching	
Optimizer	Adam
Learning Rate	8e-5
Training epochs	1000

Table 2. The experimental details on FASHION-MNIST dataset.

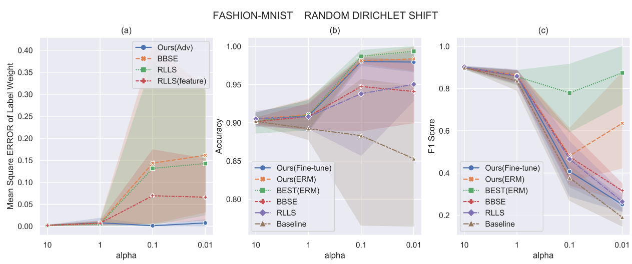


Figure 1. (a) Mean squared errors of estimated label weights (Lower is better), (b) accuracy and (c) F-1 score (Higher is better) on FASHION-MNIST for the uniform training set and random Dirichlet shifted test set, where the smaller *alpha* corresponds to the bigger shift.

Classifier Details	
Architecture	Two-layer Network
Batch Size	128
Training epochs	20
Optimizer	SGD
Learning Rate	1e-2
L2 Penalty Parameter	5e-4
Label Transformation Details	
Architecture	One-Layer Network
Label Influence Recovery Details	
Generator Architecture	Four-layer Network
Training Method	TAC-GAN (Gong et al., 2019)
Distribution Matching	
Optimizer	Adam
Learning Rate	8e-5
Training epochs	1000

Table 3. The experimental details on MNIST dataset.

Figure 2. (a) Mean squared errors of estimated label weights (Lower is better), (b) accuracy and (c) F-1 score (Higher is better) on FASHION-MNIST for the uniform training set and Tweak-One shifted test set, where *alpha* is the probability of the tweaked class.

Figure 3. (a) Mean squared errors of estimated label weights (Lower is better), (b) accuracy and (c) F-1 score (Higher is better) on FASHION-MNIST for the uniform training set and minority-class shifted test set, where *alpha* is the ratio of minority classes.

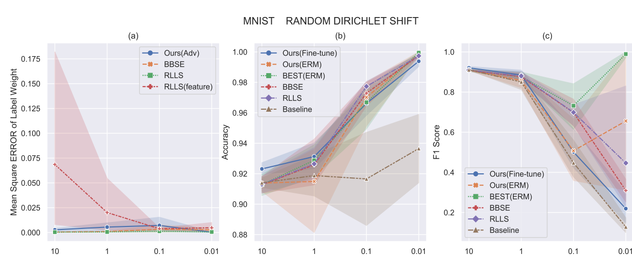


Figure 4. (a) Mean squared errors of estimated label weights (Lower is better), (b) accuracy and (c) F-1 score (Higher is better) on MNIST for the uniform training set and the random Dirichlet shifted test set, where the smaller *alpha* corresponds to the bigger shift.

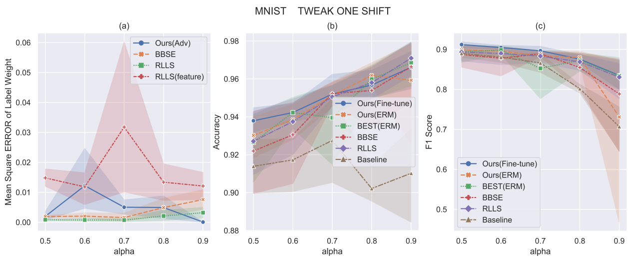


Figure 5. (a) Mean squared errors of estimated label weights (Lower is better), (b) accuracy and (c) F-1 score (Higher is better) on MNIST for the uniform training set and Tweak-One shifted test set, where *alpha* is the probability of the tweaked class.

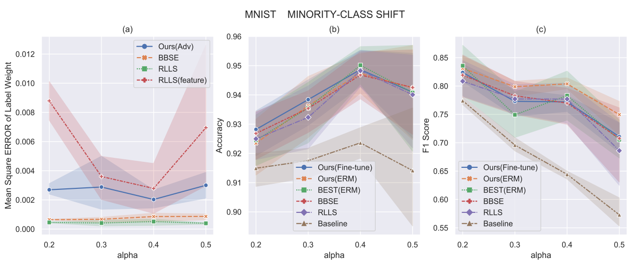


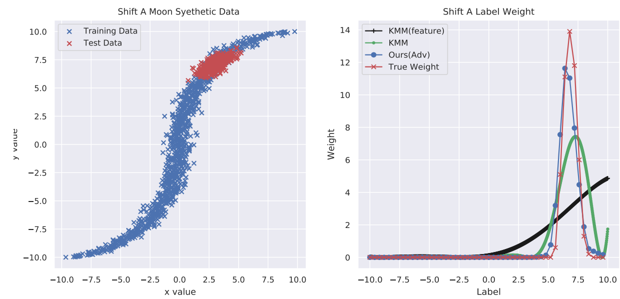
Figure 6. (a) Mean squared errors of estimated label weights (Lower is better), (b) accuracy and (c) F-1 score on MNIST for the uniform training set and minority-class shifted test set, where *alpha* is the ratio of minority classes.

A.3. Label Weights Visualization of Continuous Synthetic Data Experimens

Regressor Details	
Architecture	Three-layer Network
Batch Size	64
Training epochs	1000
Optimizer	Adam
Learning Rate	1e-3
Label Transformation Details	
Architecture	Three-layer Network
Label Influence Recovery Details	
Generator Architecture	Three-layer Network
Training Method	TAC-GAN (Gong et al., 2019)
Distribution Matching	
Optimizer	Adam
Learning Rate	1e-3
Training epochs	10000

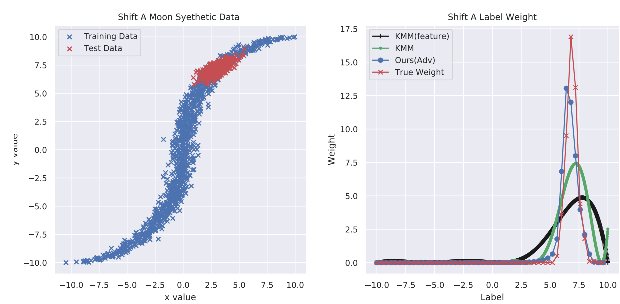
Table 4. The experimental details on Moon Synthetic dataset.

A.3.1. RESULTS OF SHIFT A



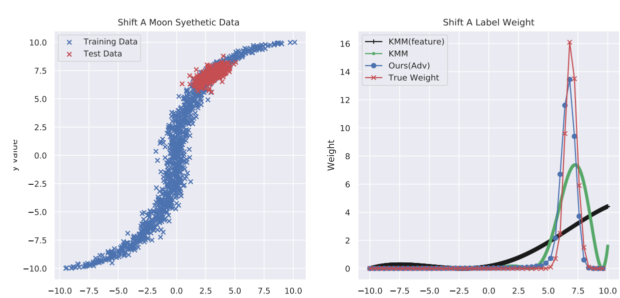
(a) (b)

Figure 7. (a) The illustration of Moon Synthetic Data (Shift A, 1st experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



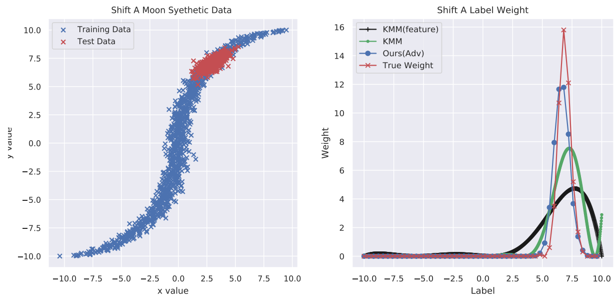
(a) (b)

Figure 8. (a) The illustration of Moon Synthetic Data (Shift A, 2nd experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



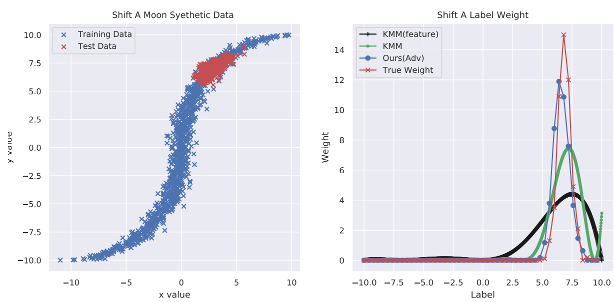
(a) (b)

Figure 9. (a) The illustration of Moon Synthetic Data (Shift A, 3rd experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



(a) (b)

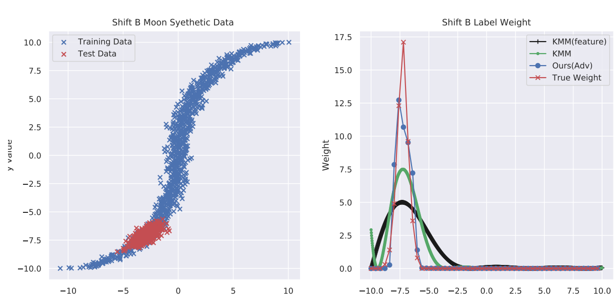
Figure 10. (a) The illustration of Moon Synthetic Data (Shift A, 4th experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



(a) (b)

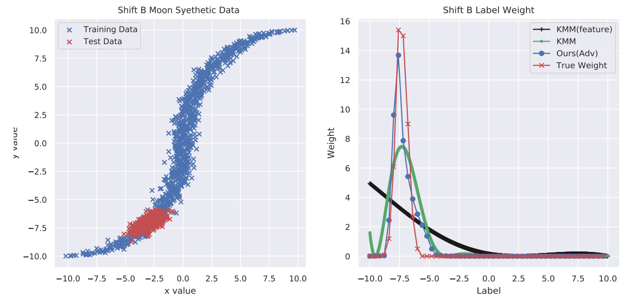
Figure 11. (a) The illustration of Moon Synthetic Data (Shift A, 5th experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.

A.3.2. Results of Shift B



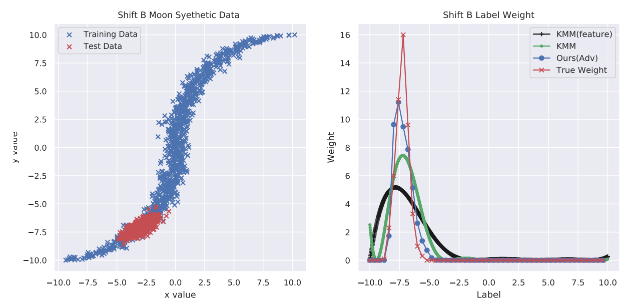
(a) (b)

Figure 12. (a) The illustration of Moon Synthetic Data (Shift B, 1st experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



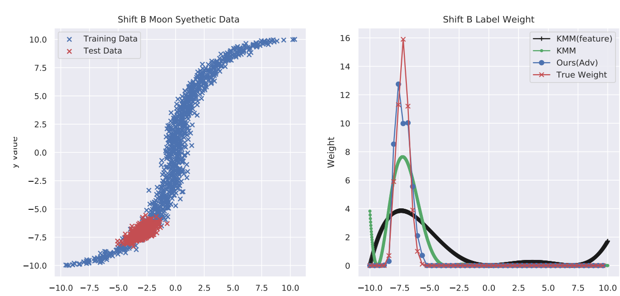
(a) (b)

Figure 13. (a) The illustration of Moon Synthetic Data (Shift B, 2nd experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



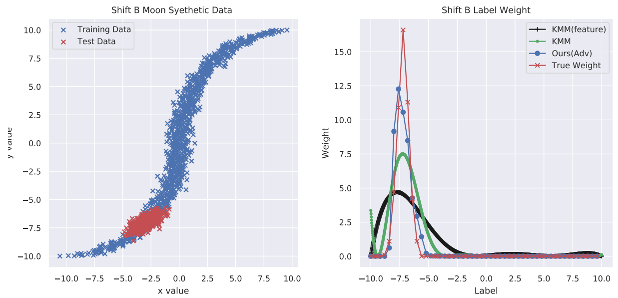
(a) (b)

Figure 14. (a) The illustration of Moon Synthetic Data (Shift B, 3rd experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



(a) (b)

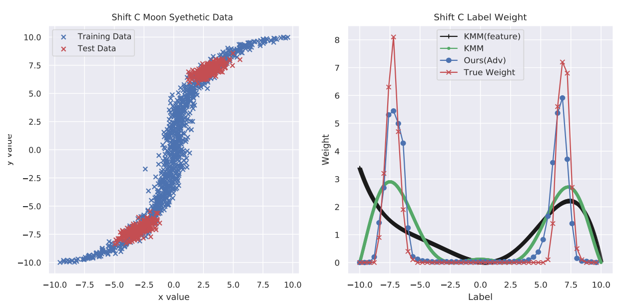
Figure 15. (a) The illustration of Moon Synthetic Data (Shift B, 4th experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



(a) (b)

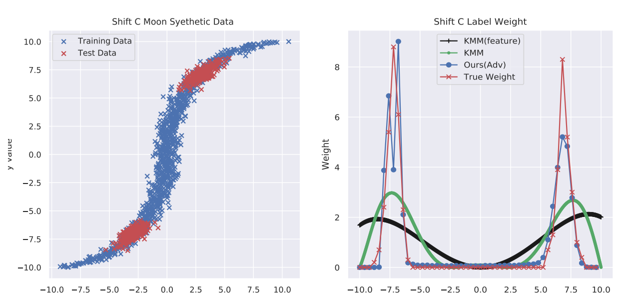
Figure 16. (a) The illustration of Moon Synthetic Data (Shift B, 5th experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.

A.3.3. Results of Shift C



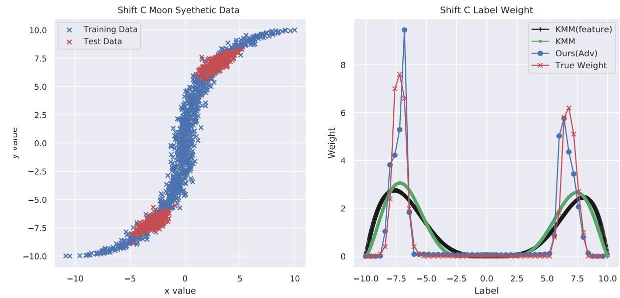
(a) (b)

Figure 17. (a) The illustration of Moon Synthetic Data (Shift C, 1st experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



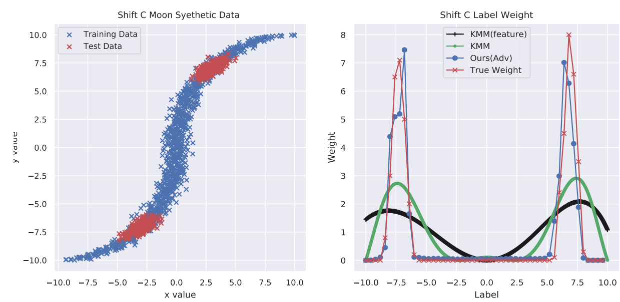
(a) (b)

Figure 18. (a) The illustration of Moon Synthetic Data (Shift C, 2nd experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



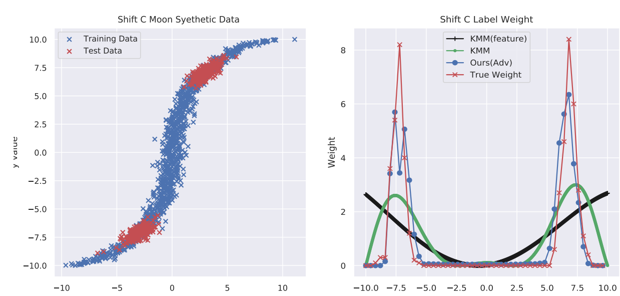
(a) (b)

Figure 19. (a) The illustration of Moon Synthetic Data (Shift C, 3rd experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



(a) (b)

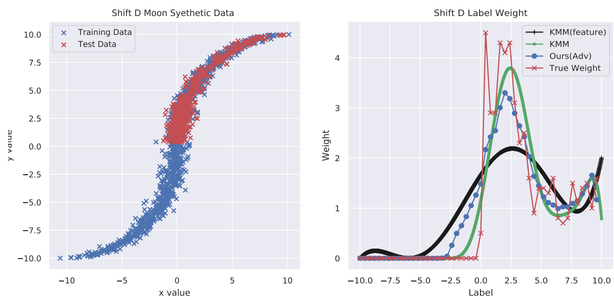
Figure 20. (a) The illustration of Moon Synthetic Data (Shift C, 4th experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



(a) (b)

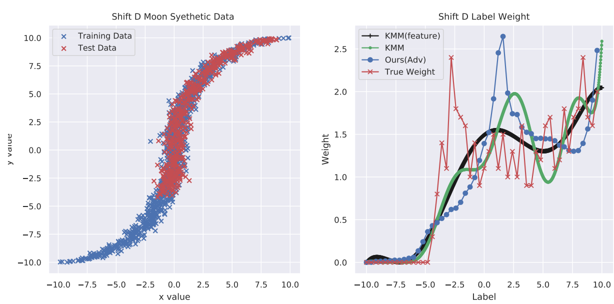
Figure 21. (a) The illustration of Moon Synthetic Data (Shift C, 5th experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.

A.3.4. RESULTS OF SHIFT D



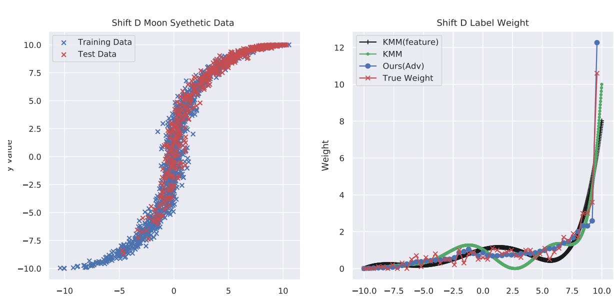
(a) (b)

Figure 22. (a) The illustration of Moon Synthetic Data (Shift D, 1st experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



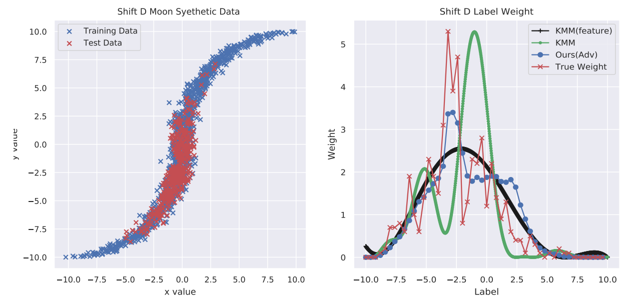
(a) (b)

Figure 23. (a) The illustration of Moon Synthetic Data (Shift D, 2nd experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



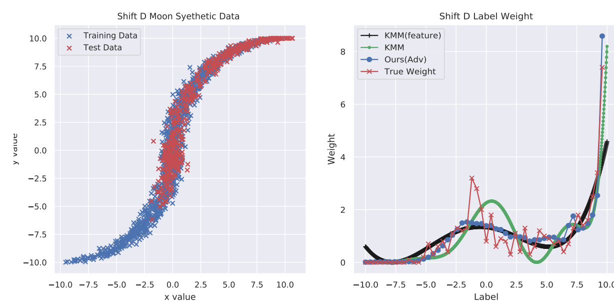
(a) (b)

Figure 24. (a) The illustration of Moon Synthetic Data (Shift D, 3rd experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



(a) (b)

Figure 25. (a) The illustration of Moon Synthetic Data (Shift D, 4th experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.



(a) (b)

Figure 26. (a) The illustration of Moon Synthetic Data (Shift D, 5th experiment), (b) The visualization of label weight P_Y^T/P_Y^S of KMM, KMM(feature), our framework and the Ground Truth.

A.4. Results of dsprite Dataset

The details of experimental settings of dsprite could be found at the table 5. The results of it could be found at the main paper.

Regressor Details	
Architecture	Discriminator of DCGAN
Batch Size	128
Training epochs	500
Optimizer	Adam
Learning Rate	1e-4
Label Transformation Details	
Architecture	Three-Layer Network
Label Influence Recovery Details	
Generator Architecture	Generator of DCGAN
Training Method	TAC-GAN (Gong et al., 2019)
Distribution Matching	
Optimizer	Adam
Learning Rate	5e-5
Training epochs	2000

Table 5. The experimental details on dsprite dataset.

References

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Mingming Gong, Yanwu Xu, Chunyuan Li, Kun Zhang, and Kayhan Batmanghelich. Twin auxiliary classifiers gan. *arXiv preprint arXiv:1907.02690*, 2019.