Table 1: Average of AUCs [%] over all target domains of each dataset.

Dataset	ProT	ProS	AEAUC	ProT($\lambda = 0$)	ProS($\lambda = 0$)	AEAUC($\lambda = 0$)	ProT($\beta = 0$)	ProS($\beta = 0$)
MNIST-r	**96.6(3.2)**	96.0(3.6)	94.3(4.9)	70.9(4.4)	64.1(4.8)	63.2(4.9)	95.8(4.0)	95.5(4.2)
Anuran Calls	**99.8(0.6)**	96.8(3.8)	96.9(4.0)	71.3(15.)	28.8(16.)	28.1(16.)	**97.9(8.7)**	91.9(11.)
Landmine	**72.4(11.)**	**72.4(11.)**	69.1(9.4)	55.3(8.1)	55.4(9.3)	54.4(8.8)	**72.1(11.)**	**71.8(11.)**
IoT	**98.4(1.7)**	**98.5(1.6)**	97.9(3.1)	84.6(8.5)	79.9(9.9)	77.3(13.)	**98.5(1.4)**	**98.5(1.5)**

We would like to thank the reviewers for their feedback and insightful comments, which we shall address below.

Reviewer #1

>For MNISR-r, the author...: As you mentioned, the domains of MNIST-r correspond to 15-degree rotation increments, which is described in the supplemental material (L12-13). We will specify this in the revised main paper. **5. Improvements 1):** We will give some intuition about our approach. Our method might not be effective when the source and target domains become less related. However, our method would reduce the negative effects of this irrelevance since it considers the uncertainty of the latent domain vectors. That is, when the distributions of the source and target normal instances differ greatly or there is a small amount of target instances, the latent domain vector of the target domain would have large variance $\sigma^2_\phi(\mathbf{X}^-_d)$. This variance alleviates the negative transfer since it prevents over-fitting. When target normal instances are available for training, our method would future reduce the negative transfer since the scores of target normal instances are directly learned to become low. We will investigate this. **2):** We will add the detailed explanation of the computational complexity for the inference described in Sec. 4.4. To infer the parameters $\mu_{\phi_*}(\mathbf{X}^-_{d'})$ and $\ln \sigma_{\phi_*}(\mathbf{X}^-_{d'})$ from $\mathbf{X}^-_{d'}$, our method requires $N^-_{d'}$ feed-forward passes of the instances in the neural networks with the parameter ϕ_*. Besides, to sample L latent domain vectors, our method requires L samplings from the standard Gaussian distribution. Note that sampling from the standard Gaussian distribution is lightweight. As a result, the total computation complexity for the inference of the anomaly score function $s(\cdot)$ becomes $\mathcal{O}(N^-_{d'} + L)$. We will clarify this in the revised paper. We evaluated the inference time of ProS on IoT. The experimental setup is the same as that in Sec. 3.5 of the supplemental material. The inference time of the $s(\cdot)$ with $L = 10$ was 5708 times faster than the training time of ProT for 100 epochs (0.0025 sec. vs 14.27 sec.) without losing the detection performance. This result shows the benefit of not needing to retrain. **3):** Thank you for your suggestion. We will add the pseudo-code in the revised paper.

Reviewer #2

2. Detailed comments (1): Our method does not require the exact number of normal/anomalous instances. Our method can use unlabeled instances, which are unknown whether anomalous or normal, by treating them as normal instances assuming most of the unlabeled instances are normal. This technique is commonly used in unsupervised anomaly detection methods [6,8]. **(2):** Our method would reduce the harmful effects of the data size difference of each domain since the objective function for each domain Eq. (3) is normalized by the data size. As to the neural networks for the latent domain vectors, we can also reduce these effects by taking average of $\eta(\mathbf{x}_{dn})$ as described in the supplemental material (L64-68). Indeed, our method with this architecture worked well against the imbalanced size datasets (Anuran Calls and Landmine) in the experiment. **(3):** Our method can detect anomalies that are partially overlapped with normal instances when a large value is set to the regularization parameter λ in Eq. (3) since training anomalies are well learned as the larger value λ. However, in this case, there is a risk that false detection of overlapped normal instances increases. In practice, we would select the appropriate value of λ using validation data on the source domains.

Reviewer #3

1. Contributions 1): Although we used AEs as the anomaly score functions due to their simplicity and effectiveness, which have been described in many studies [6,43,56,12,1], the proposed framework can use other semi-supervised anomaly detection methods with learnable parameters such as autoregressive models [19,46] and flow-based models [15,31] as described in Introduction and Sec. 4 (L62-66 and L159-161). Thus, our framework is not limited to the performance of AEs. **5. Improvements 1):** Our method can be applied to CV datasets with CNN-based AEs. We will evaluate it. **2):** We used the neural network with the parameter ϕ to represent each domain (instance set) as the latent domain vector. This neural network can preserve all the properties of the instance set with suitable ρ and η as described in [54]. The parameter ϕ, which is shared among all domains, is learned so as to infer $\mathbf{z}_d$'s that can detect anomalies well in each domain by maximizing the objective function $\mathcal{L}$. The $\mathbf{z}_d$ modifies the hidden states of the original AE so that anomalies cannot be reconstructed (high anomaly score) but normal instances can (low anomaly score) in each domain. We will clarify this in the revised paper. **3):** We conducted two additional experiments. Specifically, we evaluated our method without the regularization ($\beta = 0$ in Eqs. (4) and (5)) and it without the anomalous instances ($\lambda = 0$ in Eq. (3)). Tables 1 in this response shows the results. ProT and ProS with $\lambda = 0$ obviously deteriorated the performance, which means the importance of using anomalies in the related domains. ProT and ProS with $\beta = 0$ showed similar performance with $\beta \neq 0$ in Landmine and IoT, but worse performance in MNIST-r and Anuran Calls, which suggests the efficacy of the regularization. We will add these results in the revised paper.