Replies to Reviewer6 cBZp

Q1. *In p4, you claimed that the local reconstruction is prone to overfitting, while global reconstruction is more robust. What if local reconstruction fails to reconstruct the denoised original time series? Will this happen? Will* $MSE\left(\hat{X}_L, \hat{X}_G\right)$ *perform worse than* $MSE\left(X, \hat{X}_G\right)$ *?*

Reply:

When the anomaly is very small or even close to noise, the local reconstruction $\hat{X}_L$ is not able to reconstruct the denoised original time series well, i.e., the assumption fails occasionally. However, the performance of $MSE\left(\hat{X}_L, \hat{X}_G\right)$ can still be relatively close to $MSE\left(X, \hat{X}_G\right)$ in the case of denoising failure, because $\hat{X}_L$ will appoch to X.

Q2. *In p6, line 606, you set the range of r-ratio values based on the actual anomaly proportions. This seems strange since we cannot know in advance what the ratio will be in practice. Why not follow previous work [26]? how they manually set r-ratio?*

Reply:

(1) Many baselines are sensitive to thresholds and only tested on the point-adjust scenarios. To reduce the sensitivity of the threshold in the scenario without point-adjust, we choose the real anomaly ratio as the threshold for all models' anomaly detection. The unified threshold-choosing makes the experiments concentrate on anomaly score modeling.

(2) The r-ratio is in [26] set based on the anomaly labels of the validation dataset, which is 0.1% for SWaT, 0.5% for SMD, and 1% for other datasets. However, according to the published code, SMD uses 80% of the data from the training set as the validation set, while PSM, MSL, and SMAP directly use the test set as the validation set.

W1. *This work stems from the assumption that anomalies' associations mainly concentrate on adjacent time points, which limits the applicability of their approach when assumption fails.*

Reply: Thanks for the comments. Because the unsupervised anomaly detection doesn't have any lables, a assumption for anomaly is necessary. eg. Statistical assumptions,Model-based assumptions,Distance-based assumptions,Density-based assumptions. Each assumption has its own limitation. Our assumtpion is that the anomaly is sporadicaly and manifest in specific segments that can be smoothed by global context. The noise is ubiquitous while noise that can smoothed by local context. Base on this assumption, we can use $MSE(\hat{X}_L, \hat{X}_G)$ to reduce the impact of noise, because both $\hat{X}_L$ and $\hat{X}_G$ can smoothed noise out. Because the anomaly can not be smoothed by local context but global context, we can use the $MSE(\hat{X}_L, \hat{X}_G)$ to detect anomaly.

W2. *Table 1 show a significant deviation from the results reported in existing papers.*

Reply: We statics all the data information from different works in Table a1, and found that the same dataset in different works may be different. Our datasets are close to [26], and minor differences between from GDN, TranAD, and other methods. The inconsistencies in data information are caused by diferent sources, e.g., the anomaly ratio of the MSL dataset in the published in OmniAnomaly, MTAD-GAT and AT are different.

Information statistics on data

	OmniAnomaly	GDN	MTAD-GAT	TranAD	AT	MGRD
MSL	10.72%	-	10.27%	10.72%	10.5%	10.53%
PSM	-	-	-	-	27.8%	27.76%
SMAP	13.13%	-	13.13%	13.13%	12.8%	12.79%
SMD	4.16%	-	-	4.16%	4.2%	4.21%
SWaT	-	11.97%	-	11.98%	-	12.14%
WADI	-	5.99%	-	5.99%	-	5.77%

Table a1. anomaly inform in different works

W3. *Some related baselines are missing [1,2].*

Reply: Thanks for the comments. As the time is very limited in rebuttle, we are trying to include them in our baslines. If the time and computing source is enough, we will add them in the discussion period.

W4. *Some conclusions are unfounded. In p7, line 736, (1)* $MSE(\hat{X}_L, \hat{X}_G)$ *is not always smaller than* $MSE(\hat{X}_G, X)$ *; (2)* $MSE(X, \hat{X}_L)$ *is at a large scale, which means* $\hat{X}_L$ *can't reconstruct the denoised time series effectively.*

Reply:

(1) $MSE(\hat{X}_L, \hat{X}_G)$ get slightly worse than $MSE(\hat{X}_G, X)$ on HAI and WADI. When anomaly is as small as noise, the small anomaly will be smoothed out by both of them and the $MSE\left(\hat{X}_L, \hat{X}_G\right)$ will close to 0. However, X retains the small anomaly, and therefore $MSE\left(X, \hat{X}_G\right)$ can indicate these small anomalies. The small anomaly detection is very challenging and beyond our models' ability. Utilizing $MSE(\hat{X}_L, \hat{X}_G)$ for anomaly detection can effectively reduce the impact of false alerm, because both $\hat{X}_L$ and $\hat{X}_G$ can smooth noise. To demonstrate if the MGRD can achieve fewer false positives, we report the FPR of $MSE(\hat{X}_L, \hat{X}_G)$, $MSE(X, \hat{X}_G)$, $MSE(X, \hat{X}_L)$ on all datasets in Figure a1. The lower value indicates better performance. We can see that $MSE(\hat{X}_L, \hat{X}_G)$ achieves the best on all datasets.

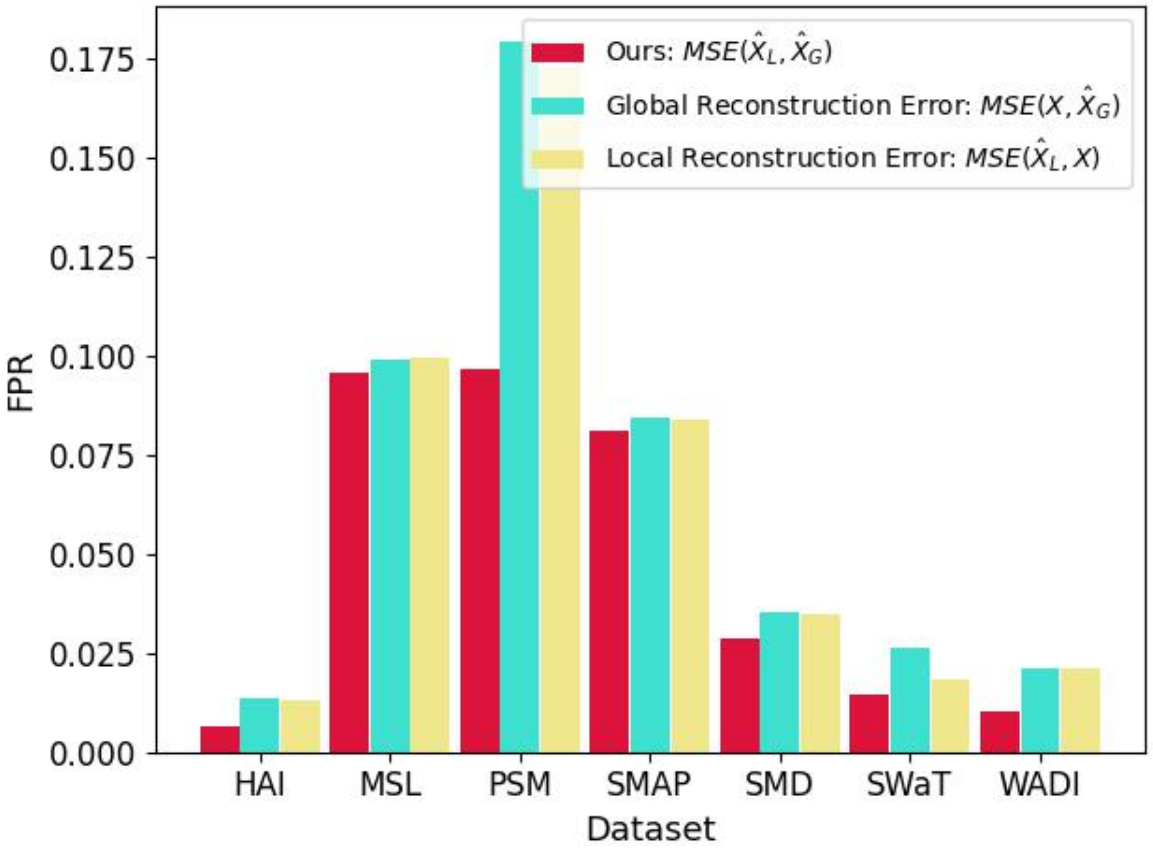

Figure a1. FPR on all datasets

(2) $\hat{X}_L$ is reconstructed by local context, and only able to smooth noise but for anomaly which manifests in segments or has large mangitude. $MSE(X, \hat{X}_L)$ is not desinged to detect anomaly. However, if there are many anomalies as small as noise, $MSE(X, \hat{X}_L)$ can get as close performance as other two MSEs shown in Figure 4 (on WADI dataset). In this case, $MSE(X, \hat{X}_G)$ may get better performance, because the small anomalies are not smoothed and kept in X. Because the noise is kept, the false positive rates of both in $MSE(X, \hat{X}_L)$ and $MSE(X, \hat{X}_G)$ are higher than that of $MSE(\hat{X}_L, \hat{X}_G)$ as shown in Figure a1.

Suggestions:

- *Only the first letter needs to be capitalized (figure and table titles).*
- *Better to plot three types of anomaly score together (Figure 7-9).*
- *Better to discuss limitations in the main paper or in the Appendix.*

Reply: Thanks for the kindly suggestions.
(1) We will revise the formats for figure and table titles.
(2) We add three types of anomaly score in Figure 7 (SMD). as shown in Figure a2. We can see that the performances of $MSE(X, \hat{X}_G)$ and $MSE(\hat{X}_L, \hat{X}_G)$ are close. However $MSE(X, \hat{X}_L)$ can not detect anomaly. This meets our basic assumption. $\hat{X}_L$ is used to smooth noise. $\hat{X}_G$ is for noise and anomaly. When noise is not very heavy, $MSE(X, \hat{X}_G)$ and $MSE(\hat{X}_L, \hat{X}_G)$ are very close. $MSE(X, \hat{X}_L)$ cannot used for anmaly detection ,because it cannot smooth anomaly.

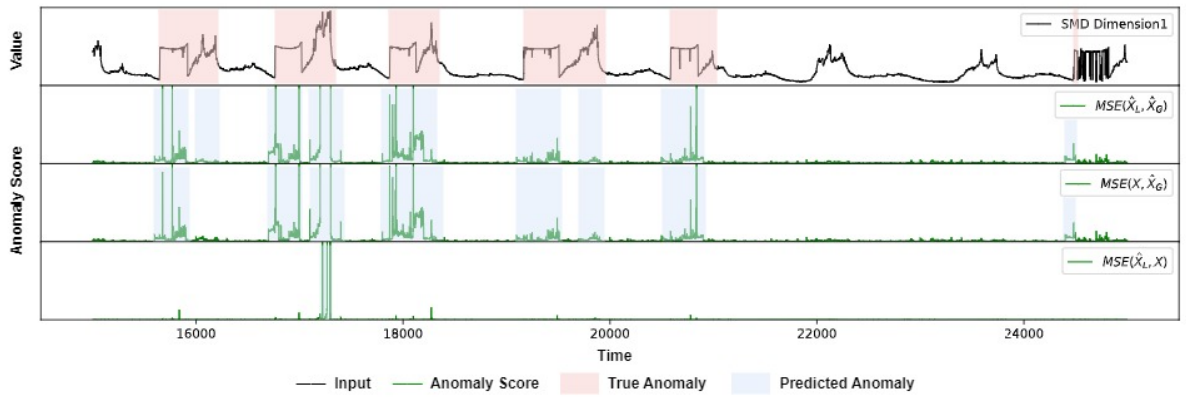

Figure a2. Three types of anomaly score on SMD

(3) We will add the limitation discussion in our main text. When the anomaly is as small as noise, both local and global reconstruction will smooth the anomaly out. In this case MGRD can not detect anomaly. This is very challenging for many models, and still open in this domain.