Replies to Reviewer q1b2

Q1. *Is the motivation of this paper to clearly distinguish noise from anomalies, or is it inspired by "the local correlations of the anomaly transformer model without explicit data reconstruction supervision may lead to trivial solutions" ?*

Reply: Thanks for the question. The motivation can be found in Lines 103-109 (Page 1) and Figure 1 (Page 2). Here, we clarify the motivation by the assumption that the **noise** (white noise) in the time series domain is ubiquitous and can be smoothed out by local reconstruction, while the **anomaly** appears within specific segments with a large impact in time and magnitude. Based on this assumption. We utilize the local reconstruction $\hat{X}_L$ to smooth noise on $\mathbf{X}$, and $\hat{X}_G$ to smooth both noise and anomaly. In the **noise area**, both $\hat{X}_L$ and $\hat{X}_G$ can smooth the noise and get very similar reconstructions. In the **anomaly area**, only $\hat{X}_G$ can smooth out the abnormal change by remote context, and therefore $|\hat{X}_L - \hat{X}_G|$ will be enlarged to indicate the anomaly.

The trivial solution of the anomaly transformer is that it can not utilize its assumption of Association Discrepancy to detect anomalies. Its anomaly score $AnomalyScore(X) = Softmax(-AssDis(P, S, X)) \odot |X - \hat{X}|$ needs the reconstruction error $|X - \hat{X}|$ for complement. However, our method only utilizes the reconstruction error $|\hat{X}_L - \hat{X}_G|$ as the anomaly score that is consistent with our assumption.

Q2. *In sections 3.3 and 3.4, which techniques were utilized in the local reconstruction model contributes to the improvement in the model's denoising?*

Reply: Thanks for the question. The attention matrix of transformer of $\hat{X}_L$ is added by a Gaussian kernel $G^m_{i,j} = \frac{1}{\sqrt{2\pi}\sigma^m_i} \exp\left(-\frac{|j-i|^2}{2\sigma^{m\,2}_i}\right)$, and the adjusted attention scores will enhance the influence of neighboring data in the reconstruction results. The noise point (white noise) in reconstructed result will be smoothed out by tis local neighboring points.

Q3. *Could you explain the explicit difference between using minimization optimization methods and the minimax strategy?*

Reply: Thanks for the question. Anomaly Transformer utilizes the **minimax** strategy. The **minimize** phase will make the prior-association P approach series-association S. The **maximize** phase forces points to pay more attention to the non-adjacent area.

Our method only utilizes a simple *minimization* strategy for both local- and global- reconstruction, and the regularization $KL(G^m_L||G^m_G) + KL(G^m_G||G^m_L)$.

Con1. *The overall motivation of this paper is difficult to underatnd. In Section 1, for example, "Nevertheless, the absence of explicit data reconstruction supervision in local correlations may lead to trivial solutions," however, it does not clearly demonstrate the rationale behind this motivation.*

Reply: Thanks for the comments. The clarity of motivation is illustrated in the reply to the Q1.

Con2. *The paper raises the issue in section 1 that existing models fail to explicitly distinguish noise from anomalies, but later sections do not demonstrate how this problem is effectively addressed, nor do they explain how the model emphasizes the difference between noise and anomalies.*

Reply: Thanks for the comments. We differentiate noise and anomalies by the assumption that the noise is ubiquitous white noise, while the anomaly appears sporadically and manifests within specific segments (Line 89-90, Page 1).

Our model can distinguish noise and anomaly by the following strategies:

- $|\hat{X}_L - \hat{X}_G|$ can detect anomalies.
- $|\hat{X}_L - X|$ can indicate noise points.

Con3. *The novelty of this paper is trivial. Most of the model components are derived from the one proposed in Anomaly Transformer model. Although there are some modifications, the lack of clear motivation hinders the demonstration of its innovativeness.*

Reply: Thanks for the comments. We can clarify the innovations by the following summarizations.

- **Motivation.** We considered the different between **noise** and anomaly. Anomaly Transformation only fit the anomaly by association-discrepancy.
- **Optimization.** We only utilize a single direction optimization strategy by **minimization**. Anomaly Transformation requires two direction optimization by **minmax**.
- **Anomaly Score.** We use the $|\hat{X}_L - \hat{X}_G|$ as the anomaly score which is consistent with our assumption. However, the Anomaly Transformer can not use the association-discrepancy term ($AssDis(P, S, X)$) as the anomaly score, but requires the reconstruction as the complement ($Softmax(-AssDis(P, S, X)) \odot |X - \hat{X}|$). This is not consistent with their assumption.
- **Architecture.** We utilize a dual path transformer (two outputs: $\hat{X}_L$, $\hat{X}_G$), while the Anomaly Transformation only has a single output $\hat{X}$.

Con4. *In section 3.3 and 3.4, this paper does not specifically elaborate on which part of the local modeling process in the model improves robustness to noise and cannot correspond to the conclusion.*

Reply: Thanks for the comments. As we discussed above, assuming that $\hat{X}_L$ can smooth the noise by the Transformer's local context and $\hat{X}_G$ can smooth the anomaly by global context, If we use $|\hat{X}_G - X|$ as the anomaly score, the score will be affected by noise. However, if we use $\hat{X}_L$ in the place of X then we can reduce the impact of noise. This can be illustrated by Figure 4 (Page 7).

Con5. *In section 3.5, the comparison between minimization optimization methods and minimax strategies is not explained. As one of the motivations mentioned in the paper is inspired by Anomaly Transformer model and the minimax strategy is a crucial strategy of it, the absence of an explanation for the comparison of using minimization optimization strategy, leaves the reader confused.*

Reply: Thanks for the comments. We try to detect anomalies by two reconstructions $\hat{X}_L$ and $\hat{X}_G$ which have the same optimization object of minimizing reconstruction errors. So we don't need the minmax. However, Anomaly Transformer requires minimizing the reconstruction and maximizing the association-discrepancy , and therefore applies a minmax strategy.

Con6. *There are format errors in the appendix section, e. g, it does not start on a new page.*

Reply: Thanks for the comments. We will fix it in revisions.

Reproductivity: The code is released by an anonymous URL **https://anonymous.4open.science/r/MGRD**.