

# Supplementary Materials: A Multilevel Guidance-Exploration Network and Behavior-Scene Matching Method for Human Behavior Anomaly Detection

Anonymous Authors

## 1 FURTHER ELABORATION ON BSMM

As depicted in Figure 1 for clarity, we have created an explanatory diagram detailing the mechanism of the “read” process. **Notably, in experiments, we calculate similarity weights between behavior and the first  $L_b$  channels of all memory items.** We select the top  $K$  vectors with the highest similarity weights as matching weights  $C^b \in \mathbb{R}^K$  and compute scene-related anomaly scores with their corresponding scene similarity weights. In the experiments, we utilize the four datasets mentioned in Section 4.6 of the paper and split them into train, validation, and test sets with a 5:2:3 ratio. We perform hyperparameter tuning on the validation sets and select the optimal  $K$  values of 4. For this explanatory diagram, we assume a hypothetical value of  $K$  equal to 2.

Specifically, consider two query items with identical behavioral features but differing scene features, denoted as  $q_1$  and  $q_2$ . Initially, we compute the similarity between behavioral features and the first  $L_b$  channels of each item in the Behavior-scene Memory. Similarly, we calculate the similarity between scene features and the last  $L_s$  channels of each item in the Behavior-scene Memory. The results, as depicted in the figure 1 (without normalization for the illustrative figure), are then used to form matching weights  $C_b$  comprising the top  $K$  data points with the highest behavioral similarity. In the illustrative diagram, the behavioral features of the third and fourth items exhibit the highest similarity to those of the query. Next, we compute the distance between the behavior matching weights  $C^b$  and the corresponding scene similarity addressing weights  $C^s$  to derive the final scene-related anomaly score  $S_{mm}$ .

From the demonstration results, it is evident that for the query item  $q_2$ , the behavioral features are similar to the top  $K$  items in memory. However, the scene features exhibit low similarity with the last  $L_s$  channels of memory items, indicating a scene-related anomaly.

## 2 FURTHER EXPERIMENTAL DETAILS

### 2.1 Statistical Testing

We further elucidated the effectiveness of our method from a statistical perspective. To ascertain the consistency and stability of the model’s AUC performance, we conducted 20 repetitions of experiments in our approach and three open-source benchmark methods, namely JIGSAW[4], HF<sup>2</sup>-VAD[2], and STG-NF[1]. We calculated the average and standard deviation of the AUC metric for each method, as shown in the table 1, the results indicate that while the standard deviation of our algorithm is not the smallest, it still aligns with the average level. Furthermore, the average outcomes of our method in repeated experiments surpass those of other algorithms. Figure 2 further presents the distribution of AUC values from multiple experiments of our algorithm on two datasets. Due to the ShanghaiTech dataset being simpler compared to the UBnormal dataset,

Table 1: The AUC(%) of mean and standard deviation (mean±std) across 20 repetitions of experiments.

Method	ShanghaiTech	UBnormal
HF <sup>2</sup> -VAD [2]	71.2±0.7	-
JIGSAW [4]	83.2±1.2	57.1±3.5
STG-NF [1]	85.3±1.8	68.9±2.1
Ours	86.9±1.2	74.3±1.8

Table 2:  $p$ -value ( $p$ ) computed using  $t$ -tests.

Method	ShanghaiTech	UBnormal
Ours with HF <sup>2</sup> -VAD [2]	$2.1 \times 10^{-10}$	-
Ours with JIGSAW [4]	$1.8 \times 10^{-7}$	$2.1 \times 10^{-41}$
Ours with STG-NF [1]	$1.0 \times 10^{-3}$	$3.6 \times 10^{-15}$

our algorithm demonstrates a higher median, a smaller box, and shorter whiskers on the ShanghaiTech dataset, indicating superior AUC performance.

To ascertain if there is a significant difference between our method and the other three approaches, we conducted statistical calculations using  $t$ -test to calculate  $p$ -values. If the  $p$ -value is less than 0.05, it indicates a significant difference between the two algorithms. From the table 2, it is evident that the  $p$ -values for our method in comparison to the other methods are all less than 0.05, indicating a significant difference between our algorithm and baseline methods.

### 2.2 Ablation Experiments on Parameters

**LSTA module count  $L$ .** The motion features primarily focus on the actions of individuals. In the field of action recognition, Spatial-Temporal Excitation (STE)[5] is capable of effectively extracting the spatiotemporal representation of actions. We observed that incorporating a large convolutional kernel on top of STE is advantageous for capturing action anomalies. Subsequently, we investigated the impact of the LSTA module count, denoted as  $L$ , on the performance of abnormality scores’ Area Under the Curve (AUC). To control variables, we specifically calculated the AUC performance of the action anomaly detection branch.

As illustrated in Figure 3(1), the model’s performance exhibits an increasing trend, with the number of modules experiencing relatively rapid growth before reaching 5, and slowing down significantly thereafter. In order to strike a balance between the speed and performance of the model, we have determined that the optimal quantity of LSTA modules is 5.

**Masking Rate  $\gamma$ .** In this section, we analyze the impact of mask coverage  $\gamma$  in Section 3.3 of the main text on performance. It is

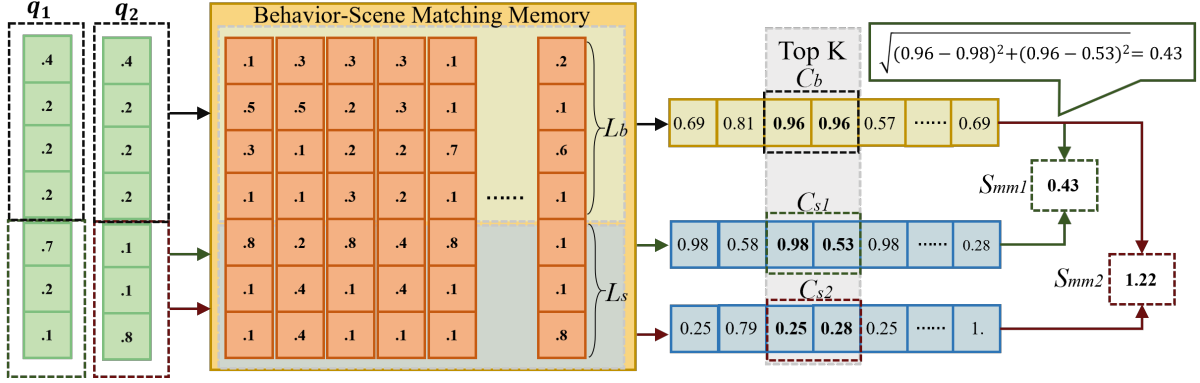


Figure 1: Flowchart of the BMSS read process and scene-related anomaly score calculation.

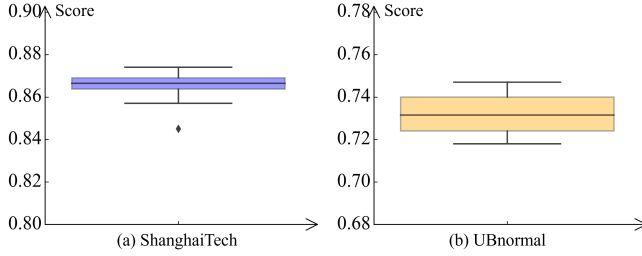


Figure 2: Our algorithm's boxplots for two datasets.

important to note that during inference, tests are grouped into  $\frac{1}{Y}$  sets, and if the count is less than two sets, they are grouped into two sets. As depicted in Figure. 3(2), within the five numerical groups we defined the algorithm's AUC performance peaks at 50%. This could be attributed to the excessively small mask coverage, allowing the model to effectively learn high-level appearance features from adjacent patches. Conversely, an excessively large mask coverage hampers the normal learning of appearance features.

**$\alpha$  and  $\beta$  in Loss Function.** We conduct a two-stage fine-tuning process. Initially, we set the  $\alpha$  parameter in the loss function at intervals of 0.1 within the range of 0 to 1 to determine the optimal performance range, which was identified as 0 to 0.2. Subsequently, we fine-tuned the parameters at intervals of 0.05 to determine the final parameter settings. As depicted in Figure. 3(3), firstly, for values of  $\alpha$  below 0.1, its effect on the experiments is minor. As  $\alpha$  increases, its influence becomes more pronounced. However, once the parameter exceeds 0.1, there is a notable decline in performance. Ultimately, we select a value of 0.1 for  $\alpha$ . For the value of  $\beta$ , we employ the same value as [3], which is set to 0.0001.

**$\lambda_{app}$  in Anomaly Score.** Similarly, we employed a two-stage approach for  $\lambda_{app}$ , ultimately choosing the range of 0 to 0.2. As depicted in Figure. 3(4), the impact of  $\lambda_{app}$  on performance exhibited a similar trend to that of  $\alpha$  in the loss function, showing an initial increase followed by a decrease within the range of 0.1 to 0.2. It peaked at 0.1. Hence, the final choice for  $\lambda_{app}$  was also 0.1.

Table 3: The AUC(%) performance of Cross-dataset.

Algorithm	Test on SHT	Test on UBN
HF <sup>2</sup> -VAD [2]	74.7 ↓ 2.0	-
JIGSAW[4]	81.3 ↓ 3.6	55.7 ↓ 1.2
STG-NF[1]	84.0 ↓ 2.2	67.9 ↓ 2.6
Ours	85.1 ↓ 1.9	72.6 ↓ 2.3

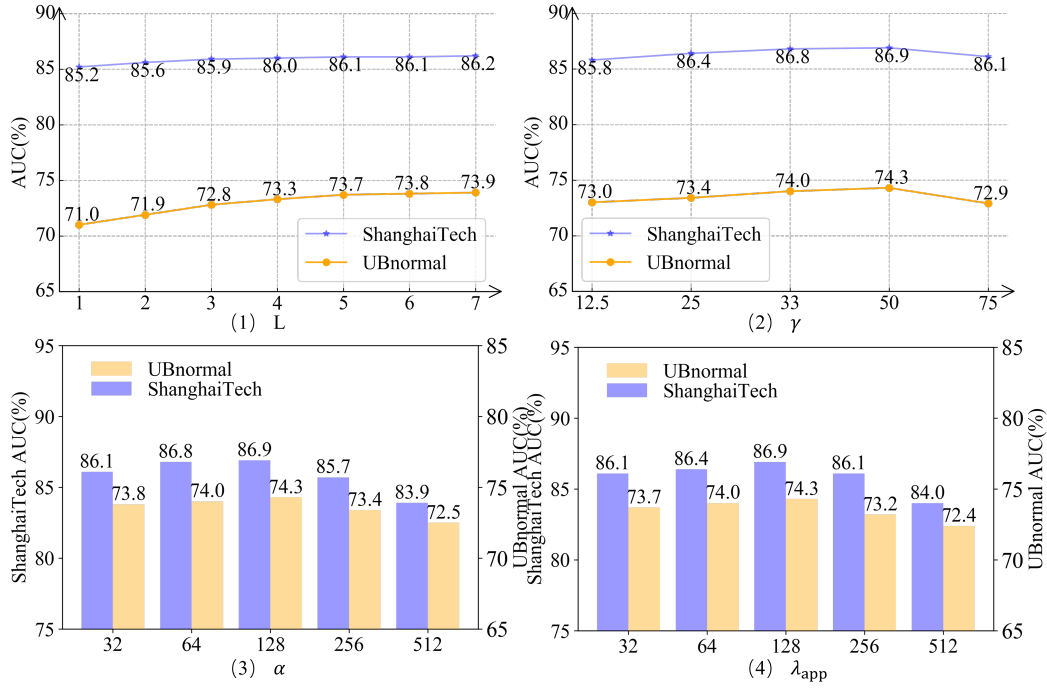
### 2.3 Cross-Dataset Generalization Capability

As shown in Table 3, our study not only addresses model performance on specific datasets but also emphasizes its broader applicability. Therefore, we employ cross-dataset validation, training on one dataset's training set and evaluating on another dataset's testing set. "Test on ShanghaiTech" referred to here implies the model is trained on the UBnormal dataset and tested on the ShanghaiTech dataset. This method provides a comprehensive insight into the model's adaptability under varied conditions. It is worth noting that, due to significant differences in the scenarios covered by the two datasets, training on one dataset to detect scene-related anomalies in another dataset is not effective. Therefore, we only discuss the robustness of the Two-level Guidance-Exploration Network. We conduct a comparison between our method and publicly available algorithms, and the experimental results reveal that our approach exhibits a slight superiority over other algorithms. Furthermore, in comparison to training and testing on the same dataset, our approach exhibits only a small degradation in performance.

## 3 DISCUSSION

The paper introduces a multi-level guided exploration network, resembling a teacher-student network in some aspects. The method takes a multimodal approach to detect different types of anomalies. Despite achieving considerable performance improvements, there are still some issues that require further discussion:

First, our Method relies on the extraction of skeleton keypoints. Therefore, if keypoints are poorly extracted for blurry video frames, it may negatively impact the performance of our action anomaly detection.



**Figure 3: The AUC(%) performance of (1) LSTA module count  $L$ , (2) masking rate  $\gamma$ , (3)  $\alpha$  in loss function and (4)  $\lambda_{app}$  in anomaly score Value on Impact.**

Second, many video anomaly detection methods, including ours, are considered video anomaly detection as an Out-of-Distribution (OOD) detection. Consequently, some unobserved normal actions are also treated as anomalies. In the future, we aim to design methods that can perceive such anomalies as normal occurrences.

Finally, scene-related anomaly detection is a worthwhile research direction. However, currently, anomalies related to scenes are complex and challenging to detect. While our method can address these issues, it is also limited. Performance may be compromised when encountering unseen scenarios. This might require extensive data or domain-specific knowledge, making it an exploration-worthy direction for anomaly detection.

## REFERENCES

- [1] Hirschorn. 2022. Normalizing Flows for Human Pose Anomaly Detection. *arXiv preprint arXiv:2211.10946* (2022).
- [2] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2021. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *CVPR*. 13588–13597.
- [3] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. 2020. Learning memory-guided normality for anomaly detection. In *CVPR*. 14372–14381.
- [4] Guodong Wang and Wang. 2022. Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In *ECCV*. Springer, 494–511.
- [5] Zhengwei Wang, Qi She, and Aljosa Smolic. 2021. Action-net: Multipath excitation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13214–13223.