Noise-Resistant Video Anomaly Detection via RGB Error-Guided Multiscale Predictive Coding and Dynamic Memory (Supplementary Material)

Han Hu[1], Wenli Du[1,2]*, Peng Liao[1], Bing Wang[1], Siyuan Fan[2]
[1] State Key Laboratory of Industrial Control Technology, ECUST, China
[2] Huzhou Institute of Industrial Control Technology, China
{han.hu, pengliao}@mail.ecust.edu.cn, {wldu, wangb07}@ecust.edu.cn, siyuanfan2001@outlook.com

A. Additional Model Details

In Table A, we provide the detailed architecture of the encoder and decoder in the RGB Error-Guided Multiscale Predictive Coding (EG-MPC) framework.

Component	Layer	fliter size	stride	Output size
Encoder	Conv1	3×3	$(2, 2)$	$128 \times 128 \times 32$
	Conv2		$(1, 1)$	$128 \times 128 \times 32$
	Conv3		$(2, 2)$	$64 \times 64 \times 64$
	Conv4		$(1, 1)$	$64 \times 64 \times 64$
Decoder	DeConv1		$(1, 1)$	$64 \times 64 \times 64$
	DeConv2		$(2, 2)$	$128 \times 128 \times 32$
	DeConv3		$(1, 1)$	$128 \times 128 \times 32$
	DeConv4		$(2, 2)$	$256 \times 256 \times 3$

Table A. The detailed architecture of the encoder and decoder in the EG-MPC framework.

B. Additional Ablation Studies and Hyperparameter Analysis

All experiments in this section are performed on the single-scene anomaly dataset Avenue [2] and the multi-scene anomaly dataset ShanghaiTech [3].

B.1. Effectiveness of the loss functions

Our method performs two-stage training for the next-frame prediction task (Pre-Task) and the predicted-frame reconstruction task (Rec-Task), respectively. Table B shows the anomaly detection performance obtained by using different combinations of loss functions during training. The results indicate that it is advantageous to jointly consider the structural similarity index measure (SSIM) [5] losses $\mathcal{L}_{\text{SSIM}}$ and $\mathcal{L}'_{\text{SSIM}}$ on the basis of the conventional prediction loss $\mathcal{L}_{\text{pre}}$ and reconstruction loss $\mathcal{L}_{\text{rec}}$. In addition, setting contrastive losses $\mathcal{L}_{\text{con}}^{\text{D}_1}$ and $\mathcal{L}_{\text{con}}^{\text{D}_2}$ for memory items in two independent dynamic memory modules (DMMs) in the reconstruction task can significantly improve the detection performance, especially for ShanghaiTech dataset where the behavior and scale of the objects are more complex. This demonstrates the effectiveness of using contrastive loss [1] to enhance the discriminability between memory items to represent diverse normal patterns.

B.2. Impact of the number N of memory items in the DMMs

To investigate the impact of the number N of memory items in the dynamic memory modules (DMMs) on detection performance, we conduct 5 sets of experiments by varying the value of N on the Avenue and ShanghaiTech datasets. The specific results are displayed in Table C. Obviously, for Avenue and ShanghaiTech, the most appropriate values of N are 300 and 500, respectively. In contrast, either smaller or larger values of N lead to suboptimal performance in anomaly detection. This is because when there are fewer memory items, the diverse normal patterns are difficult to be expressed effectively, which leads to an increase in the reconstruction error of normal frames and makes it prone to the false alarm problem. When there are too many memory items, the probability of abnormal frames (especially for the early stages of abnormal occurrence) being well reconstructed increases, and thus their reconstruction error decreases, which can lead to miss alarms. Therefore, it is necessary to determine an appropriate N value based on the size and complexity of the dataset to ensure good detection performance.

B.3. Impact of the threshold λ_w in the sparse aggregation strategy

In the process of reconstructing each query, we introduce a threshold λ_w to sparsify the base aggregation weights corresponding to each memory item, thus increasing the difficulty of reconstructing abnormal frames. To explore the impact of the threshold λ_w on the detection performance, we conduct 8 sets of experiments by varying the value of λ_w on the Avenue and ShanghaiTech datasets (Note that for

*Corresponding author.

Avenue, $N = 300$ and for ShanghaiTech, $N = 500$.). The results are presented in Table D. For the experiments on two datasets, we notice that setting λ_w to $1/N$ resulted in the best detection performance. When λ_w is set to $3/4N$, $5/4N$, or $3/2N$, the detection performance is still relatively ideal with a small decrease. However, when the value of λ_w is further reduced (*i.e.*, $\lambda_w = 1/4N$ or $1/2N$), the base aggregation weights are less sparsified, which may allow abnormal frames to be well reconstructed thus leading to miss alarms. When the value of λ_w is further increased (*i.e.*, $\lambda_w = 7/4N$ or $2/N$), the sparsification of the base aggregation weights is relatively high, which tends to increase the reconstruction error of the normal frames and thus leads to false alarms. Overall, it is important to set an appropriate value of λ_w in the sparse aggregation strategy of memory items to achieve high performance video anomaly detection.

B.4. Impact of S_{self} and S_{ba} in the triggering conditions for the update of memory items during the testing phase

During the testing phase, we set two triggering conditions for selective update of memory items in the DMMs to reduce the chance of abnormal patterns being recorded. Specifically, after a video inference is completed, we select frames to update the memory items that meet the following two conditions: (1) its anomaly score is lower than S_{self}, and (2) the anomaly scores of the 20 frames before and after it are lower than S_{ba}. To explore the impact of the anomaly score thresholds S_{self} and S_{ba} on the detection performance, we perform 64 sets of experiments on the Avenue and ShanghaiTech datasets by combining different values of S_{self} and S_{ba}. The visualization results are shown in Fig. A and Fig. B. For the experiments on two datasets, we note that setting S_{self} to 0.15 and S_{ba} to 0.30 resulted in the best detection performance. Meanwhile, our method still demonstrates good detection performance when the values of S_{self} and S_{ba} are slightly altered. However, when the value of either S_{self} or S_{ba} is too small or too large, the performance of anomaly detection decreases significantly. This is because when S_{self} or S_{ba} is too small, there are few frames selected from the test video for the updating of memory items, which prevents the dynamic memory module from recording diverse normal patterns. When S_{self} or S_{ba} is too large, it may result in some abnormal frames being selected for updating the memory items, which inevitably has a negative impact on anomaly detection. Therefore, for the memory item updating process in the testing phase, it is necessary to set appropriate score thresholds S_{self} and S_{ba} for the triggering conditions to achieve high-performance anomaly detection.

C. Additional Visualization Results

Figure C, Figure D and Figure E show the anomaly detection results of our method on test instances from the Avenue [2], ShanghaiTech [3] and UCF-Crime [4] datasets, respectively. Firstly, through the presented anomaly score curves we can observe that:

- Our method is able to respond quickly to various anomalies, and the high score regions match well with the ground truth anomalies, demonstrating excellent AUC performance.
- For both normal and abnormal parts of the video, our method can yield relatively stable anomaly scores with significant divergence, which indicates that it has good discriminative ability.

In addition, we visualize the reconstruction RGB error maps corresponding to some of the ground truths in the figures, and it can be noted that:

- Our method has noticeable reconstruction errors in the regions where the anomalies occur.
- Our method has a small reconstruction error for normal background regions, which indicates that it can effectively overcome the interference of background noise. Especially when dealing with challenging instances in the ShanghaiTech and UCF-Crime datasets, our method demonstrates good robustness to complex scene factors such as small-scale objects, crowded crowds, and dim light intensity.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[2] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 1, 2, 5

[3] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*, pages 341–349, 2017. 1, 2, 6

[4] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 2, 7

[5] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

ID	Pre-Task		Rec-Task				Avenue		SHT	
	$\mathcal{L}_{\text{pre}}$	$\mathcal{L}_{\text{SSIM}}$	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}'_{\text{SSIM}}$	$\mathcal{L}^{\text{D}_1}_{\text{con}}$	$\mathcal{L}^{\text{D}_2}_{\text{con}}$	AUC	ΔS	AUC	ΔS
1	✓	-	✓	-	-	-	90.3	0.343	81.1	0.229
2	✓	✓	✓	✓	-	-	91.2	0.362	82.7	0.265
3	✓	✓	✓	✓	✓	✓	**92.9**	**0.431**	**86.0**	**0.349**

Table B. The AUC(%) and ΔS performance obtained by using different combinations of loss functions during training on the Avenue and ShanghaiTech datasets. The best performing results are marked in bold and highlighted.

Dataset	Avenue					SHT				
N	100	200	300	400	500	300	400	500	600	700
AUC	91.26	92.35	**92.91**	92.84	92.70	85.07	85.69	**86.02**	85.93	85.81
ΔS	0.351	0.386	**0.431**	0.417	0.398	0.296	0.328	**0.349**	0.344	0.331

Table C. The variation of AUC (%) and ΔS performance with respect to the number N of memory items in the dynamic memory modules (DMMs) on the Avenue and ShanghaiTech datasets. The best performing results are marked in bold and highlighted.

λ_w		$1/4N$	$1/2N$	$3/4N$	$1/N$	$5/4N$	$3/2N$	$7/4N$	$2/N$
Avenue	AUC	89.7	91.4	92.4	**92.9**	92.7	92.3	91.6	90.8
	ΔS	0.303	0.364	0.410	**0.431**	0.418	0.397	0.371	0.334
SHT	AUC	82.3	84.0	85.2	**86.0**	85.9	85.6	84.9	84.1
	ΔS	0.248	0.295	0.326	**0.349**	0.337	0.322	0.301	0.274

Table D. The variation of AUC (%) and ΔS performance with respect to the threshold λ_w in the sparse aggregation strategy on the Avenue and ShanghaiTech datasets (Note that for Avenue, $N = 300$ and for ShanghaiTech, $N = 500$.). The best performing results are marked in bold and highlighted.

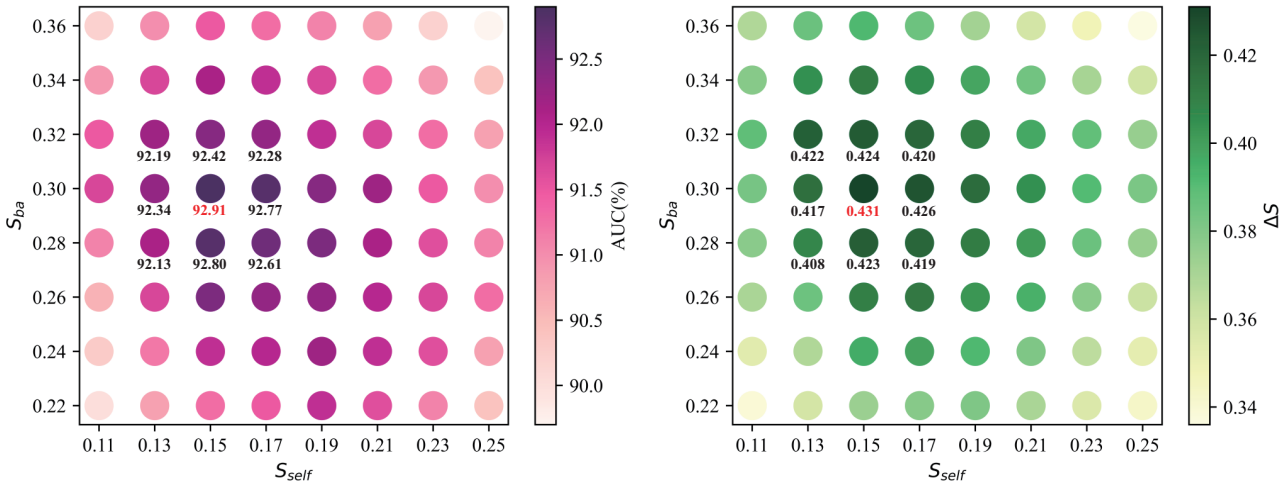

Figure A. The AUC (%) (left) and ΔS (right) performance corresponding to different combinations of score thresholds S_{self} and S_{ba} on the Avenue dataset. Specifically, after a video inference is completed, we select frames to update the memory items that meet the following two conditions: (1) its anomaly score is lower than S_{self}, and (2) the anomaly scores of the 20 frames before and after it are lower than S_{ba}. Best viewed in color.

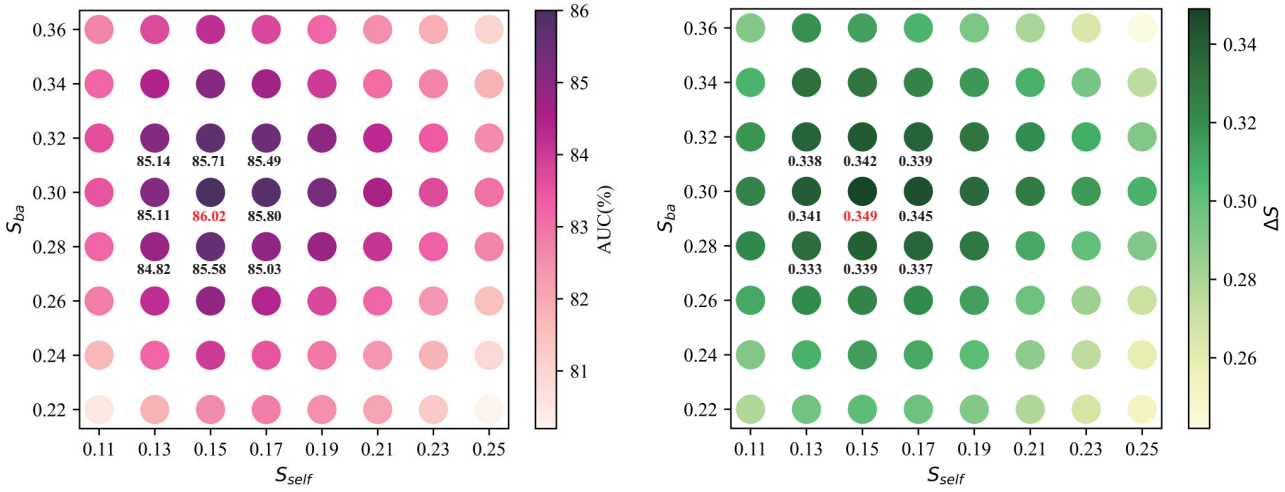

Figure B. The AUC (%) (left) and ΔS (right) performance corresponding to different combinations of score thresholds S_{self} and S_{ba} on the ShanghaiTech dataset. Best viewed in color.

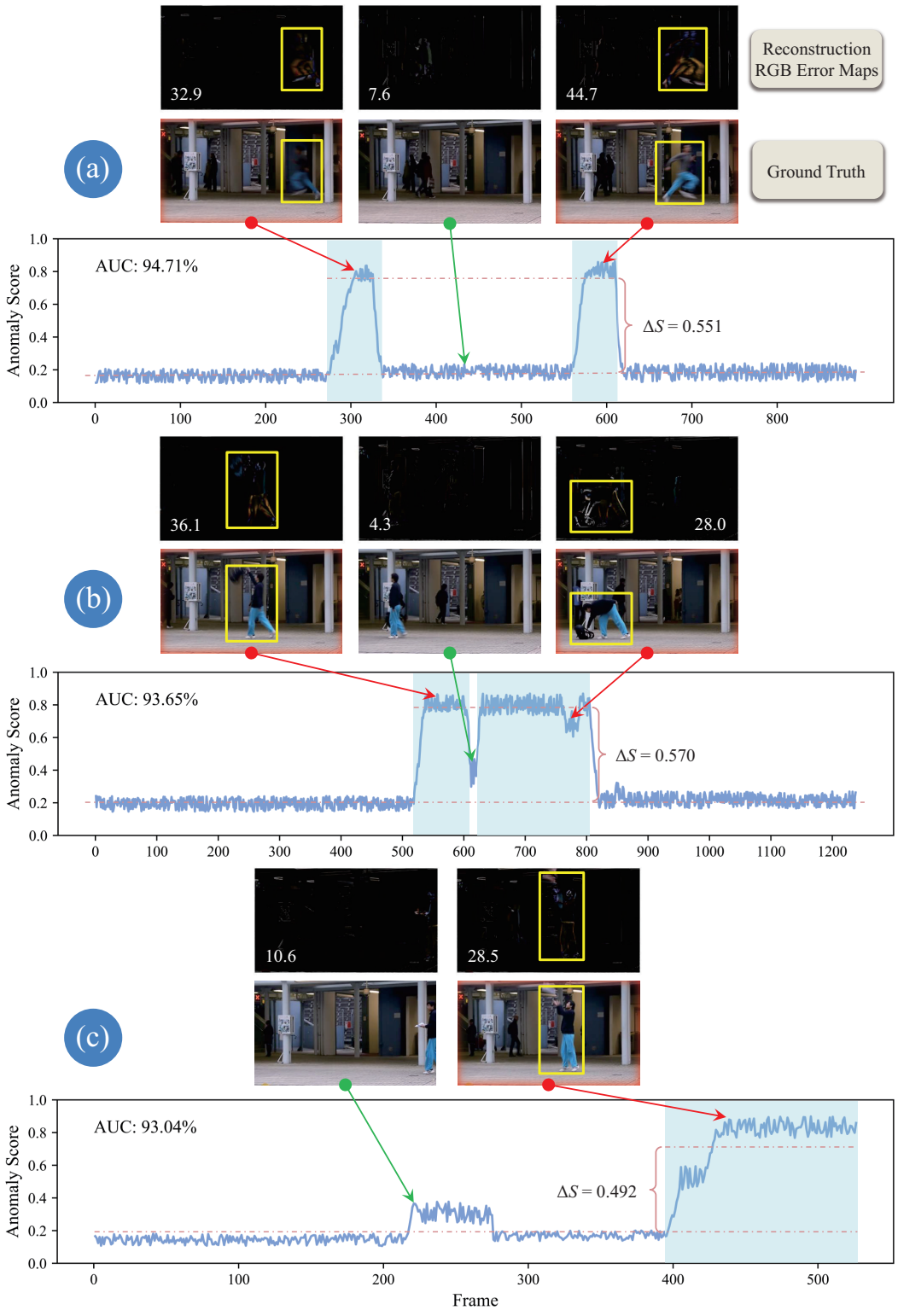

Figure C. Anomaly detection results of our method on test videos (a) 03, (b) 12 and (c) 13 of the Avenue [2] dataset. For each subfigure (*i.e.*, (a), (b), and (c)), the reconstruction RGB error maps (where the number denote the sum-square-error of the reconstructed frame compared to the ground truth), the ground truth (where the abnormal region is marked by the yellow bounding box), and the anomaly score curve (where the blue highlights represent the true anomalies) are displayed from top to bottom. Best viewed in color. Please enlarge the PDF for clarity.

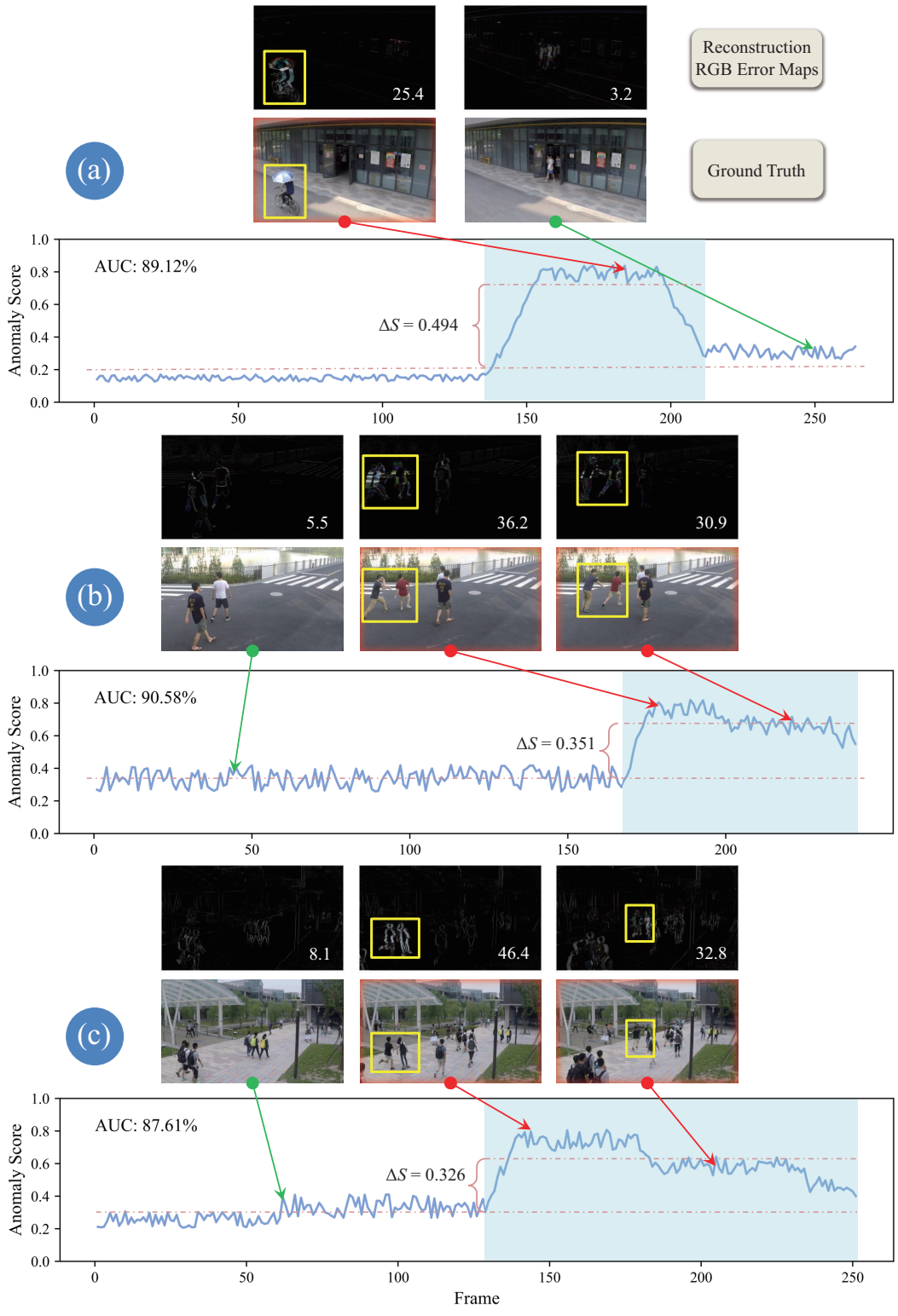

Figure D. Anomaly detection results of our method on test videos (a) 06_0147, (b) 07_0048 and (c) 08_0044 of the ShanghaiTech [3] dataset. For each subfigure (*i.e.*, (a), (b), and (c)), the reconstruction RGB error maps (where the number denote the sum-square-error of the reconstructed frame compared to the ground truth), the ground truth (where the abnormal region is marked by the yellow bounding box), and the anomaly score curve (where the blue highlights represent the true anomalies) are displayed from top to bottom. Best viewed in color. Please enlarge the PDF for clarity.

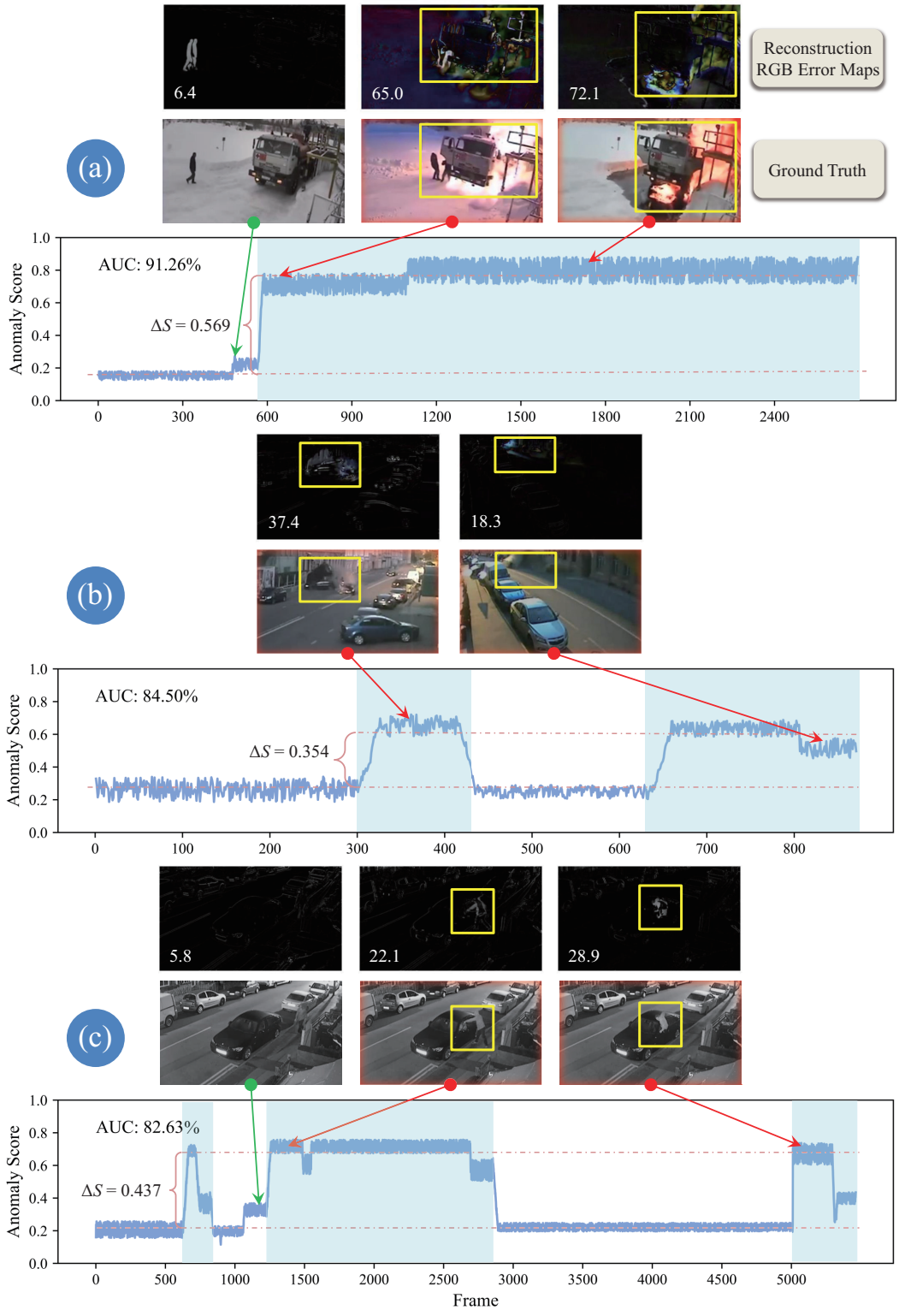

Figure E. Anomaly detection results of our method on test videos (a) 0003, (b) 0004 and (c) 0012 of the UCF-Crime [4] dataset. For each subfigure (*i.e.*, (a), (b), and (c)), the reconstruction RGB error maps (where the number denote the sum-square-error of the reconstructed frame compared to the ground truth), the ground truth (where the abnormal region is marked by the yellow bounding box), and the anomaly score curve (where the blue highlights represent the true anomalies) are displayed from top to bottom. Best viewed in color. Please enlarge the PDF for clarity.