

Unified Video Anomaly Detection Model for Detecting Different Anomaly Types

Supplementary Material

In this supplement, we provide the followings:

- Experiment 6 for effect of the prediction in visual streams.
- Experiment 7 for effect of the time T of objects or frames.
- Experiment 8 for effect of α , β , and γ in inference.
- More visual results 9 for skeleton, local-visual, and global-visual streams.
- Experiment 10 for visual feature extractors.
- Running time of UniVAD 11.

6. Different combinations of hyperparameters

Since the types of anomalies vary across datasets, we experiment with different combinations of α , β , and γ to find the optimal hyperparameters, as shown in Tab. 5. Overall, due to the high rate of human anomalies in all datasets, setting α to 1.00 (IDs 0-9) yields better performance across all datasets compared to other settings (IDs 10-21). In the NWPU dataset, where capturing scene-dependency is crucial, ID 0 demonstrates significant improvements over IDs 1-2. In the ShT dataset, which contains numerous human anomalies and a few appearance anomalies, ID 5 outperforms IDs 2 and 8 significantly. Similarly, in the UB dataset, which includes numerous human anomalies and few nonobject and appearance anomalies, ID 8 achieves significant gains compared to IDs 6-7.

7. Effect of the prediction in visual streams

Tab. 6 shows the results of the ablation study conducted on the ShT and UB datasets to demonstrate that predicting past, present, and future features effectively improves performance for visual streams compared to predicting other features. Predicting key frame features (past, present, and future) efficiently trains the model better than predicting all features, resulting in improved detection performance.

8. Effect of the time of objects or frames

In Tab. 7, we conduct an ablation study on the ShT and UB datasets to analyze the impact of the time of objects or frames on detection performance. In general, as the time T increases, prediction becomes more challenging, leading to decreased performance. In the skeleton stream, the optimal T values are 24 and 16 for the ShT and UB datasets, respectively. In the local-visual and global-visual streams, the optimal T value is 8 for both the ShT and UB datasets.

ID	α	β	γ	ShT	UB	NWPU
0	1.00	1.00	1.00	86.3	70.1	73.4
1	1.00	1.00	0.10	88.5	69.3	71.3
2	1.00	1.00	0.01	88.4	68.7	70.2
3	1.00	0.10	1.00	78.8	71.3	72.2
4	1.00	0.10	0.10	89.3	79.3	72.2
5	1.00	0.10	0.01	89.5	78.6	69.5
6	1.00	0.01	1.00	75.9	71.1	71.9
7	1.00	0.01	0.10	86.6	81.6	71.6
8	1.00	0.01	0.01	86.7	82.7	69.0
9	1.00	0.00	0.00	85.2	81.2	65.1
10	0.10	1.00	1.00	84.4	66.7	72.5
11	0.10	1.00	0.10	86.5	65.2	70.7
12	0.10	1.00	0.01	86.3	64.4	69.5
13	0.10	0.10	1.00	73.3	64.6	70.4
14	0.10	0.01	1.00	69.1	63.7	69.8
15	0.01	1.00	1.00	84.1	66.3	72.3
16	0.01	1.00	0.10	86.1	64.6	70.5
17	0.01	1.00	0.01	85.9	63.9	69.3
18	0.01	0.10	1.00	72.5	63.4	70.1
19	0.01	0.01	1.00	68.1	62.4	69.3
20	0.00	1.00	0.00	85.9	63.7	68.8
21	0.00	0.00	1.00	67.3	62.0	69.2

Table 5. Ablation experiments of the hyperparameters α , β , and γ in terms of micro-AUC (%) on the ShT, UB, and NWPU datasets. α , β , and γ are chosen from the set [1.00, 0.10, 0.01, 0.00].

9. Visual results for three streams

Fig. 5 presents the results of anomaly detection by the skeleton, local-visual, and global-visual streams for various types of anomalies. In Fig. 5 (a), since the anomaly is a human anomaly involving running in a place where running is not allowed, the skeleton stream performs better than local-visual, global-visual stream. In Fig. 5 (b), since the anomaly is a combination of a human anomaly (a person riding a bike) and an appearance anomaly (the bike), the skeleton and local-visual stream perform good performance. In Fig. 5 (c), the anomaly is a nonobject anomaly involving smoke caused by a car accident. The skeleton and local-visual streams fail to detect the anomaly, while the global-visual stream successfully detects it.

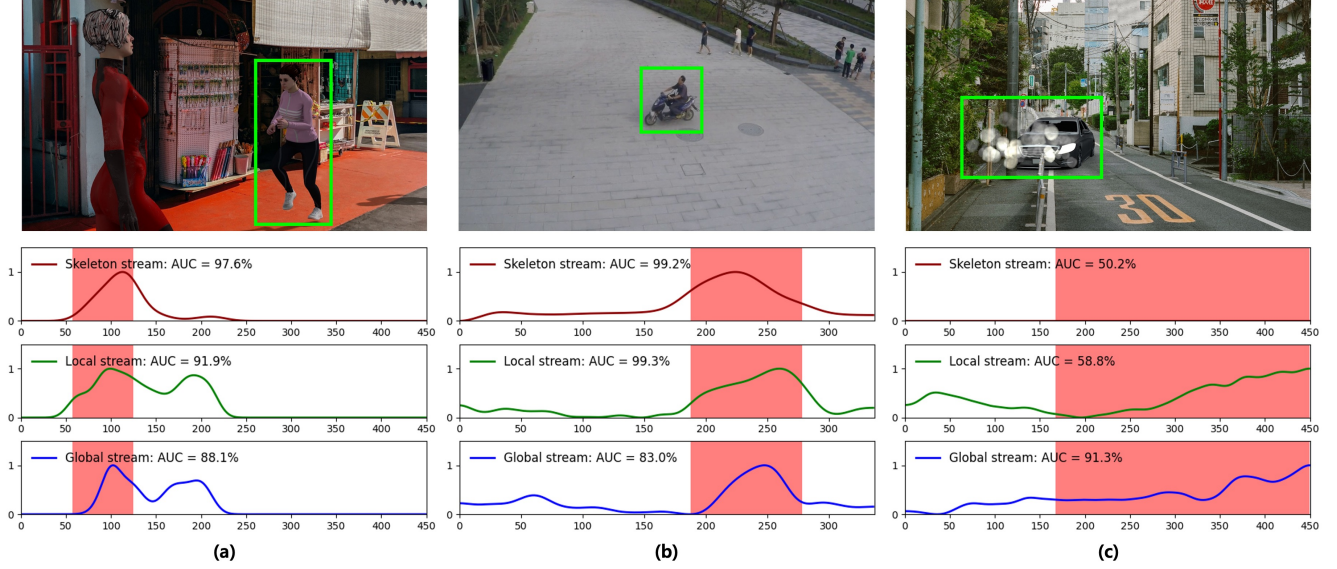


Figure 5. Comparisons of anomaly detection results for various types of anomalies across three streams. (a) The video containing human anomalies in the UB dataset. (b) The video containing appearance anomalies in the ShT dataset. (c) The video containing nonobject anomalies in the UB dataset. The metric used is micro-AUC. From top to bottom: results from the skeleton stream, results from the local-visual stream, and results from the global-visual stream.

Prediction	Local		Global	
	ShT	UB	ShT	UB
All features	82.8	62.4	52.6	58.5
Present features	83.2	58.3	62.3	61.2
Future, Past features	84.8	63.0	65.3	61.5
Future, Present, Past features	85.9	63.7	67.3	62.0

Table 6. Ablation experiments on the prediction of features in visual streams. We report the micro-AUC (%) for the ShT and UB datasets. Here, “All features” refers to predicting T features from past to future, “Present features” refers to predicting only the present features, “Future, Past features” refers to predicting both past and future features, and “Future, Present, Past features” refers to predicting past, present and future features.

T	Skeleton		Local		Global	
	ShT	UB	ShT	UB	ShT	UB
8	78.5	79.2	85.9	63.7	67.3	62.0
16	84.6	81.2	83.8	62.9	64.6	61.6
24	85.2	79.9	83.0	62.2	61.2	60.7
32	84.2	78.5	82.6	61.8	59.7	60.0

Table 7. Ablation experiments on the time T of the sequence of objects or frames in the three streams. We report the micro-AUC (%) for the ShT and UB datasets.

10. The choice of feature extractors

In Tab. 8 and Tab. 9, we compare the performance of local-, global- stream and UniVAD used with different image feature extractors. We observe, that CLIP(ViT-L-14) mostly outperforms CLIP(ViT-B-32) in experiments, but runs considerably slower. Among various CLIP versions, we selected CLIP (ViT-B/32) due to its favorable trade-off between computational efficiency and accuracy.

11. Running time

We conducted all our experiments on an NVIDIA RTX 3090 GPU. The object detection and tracker take approximately 39 milliseconds (ms) per frame. The pose extractor takes approximately 19 ms, and the visual encoder takes approximately 2.5 ms. Computing skeleton data, local-visual features, and global-visual features across all streams takes approximately 0.02 ms. UniVAD runs at 16.5 FPS with an average of 5 objects per frame.

Backbone	Local stream			Global stream			UniVAD		
	ShT	UB	NWPU	ShT	UB	NWPU	ShT	UB	NWPU
CLIP(Resnet-50)	76.8	62.7	67.9	67.8	60.9	68.0	85.7	78.3	70.6
CLIP(Resnet-101)	81.9	64.9	67.0	68.4	59.1	66.7	87.7	78.0	69.2
CLIP(ViT-B/32)	85.9	63.7	68.8	67.3	62.0	69.2	89.3	79.3	72.2
CLIP(ViT-B/16)	85.1	65.8	64.0	66.5	60.6	67.7	88.7	78.7	69.8
CLIP(ViT-L/14)	87.4	67.3	67.2	70.6	63.7	69.5	89.4	80.8	71.4

Table 8. Micro AUC-ROC (%) comparison. For each stream feature representation CLIP(ViT-B/32), CLIP(ViT-B/16), CLIP(ViT-L/14), CLIP(Resnet-50), and CLIP(Resnet-101), we mark the best scores bold.

Backbone	Local stream			Global stream			UniVAD		
	ShT	UB	NWPU	ShT	UB	NWPU	ShT	UB	NWPU
Clip(Resnet-50)	84.4	83.7	83.2	74.0	78.9	83.7	90.3	90.1	87.4
Clip(Resnet-101)	86.1	84.8	83.4	75.9	78.4	83.4	90.9	90.3	87.5
Clip(ViT-B/32)	86.8	84.8	83.8	75.1	80.0	83.6	91.5	90.1	87.6
Clip(ViT-B/16)	87.6	85.6	84.6	74.8	79.4	84.2	91.2	90.0	88.5
Clip(ViT-L/14)	89.2	85.2	85.0	75.1	82.2	84.5	91.6	91.4	88.7

Table 9. Macro AUC-ROC (%) comparison. For each stream feature representation CLIP(ViT-B/32), CLIP(ViT-B/16), CLIP(ViT-L/14), CLIP(Resnet-50), and CLIP(Resnet-101), we mark the best scores bold.