Open-Vocabulary Video Anomaly Detection

Peng Wu[1], Xuerong Zhou[1], Guansong Pang[2]*, Yujia Sun[3], Jing Liu[3], Peng Wang[1]*, Yanning Zhang[1]
[1]Northwestern Polytechnical University, [2]Singapore Management University, [3]Xidian University
{xdwupeng, zxr2333}@gmail.com, gspang@smu.edu.sg, yjsun@stu.xidian.edu.cn
neouma@163.com, {peng.wang, ynzhang}@nwpu.edu.cn

1. Data division

For OVVAD task, we divide abnormal categories into base categories and novel categories, only these samples of base category are available during the training phase. Following the most open-vocabulary learning works [1, 2], the frequent and common classes are used as base categories, while the rare classes are held out as novel categories. Specifically, on UCF-Crime, *Abuse, Assault, Burglary, Road Accident, Robbery,* and *Stealing* are designated as base categories, the rest categories fall into novel categories. On XD-Violence, *Fighting, Shooting,* and *Car accident* are considered as base. In contrast, UBnormal is used for open-set VAD, training anomalies serve as base categories, and test anomalies are classified as novel categories.

2. Diverse category analysis in NAS module

To validate the effectiveness of NAS module and simulate the real-world scenarios, we explore different categories to generate pseudo anomalies. Firstly, we randomly select half of novel categories, then use NAS module to generate pseudo anomalies; Secondly, we combine the randomly selected half of novel categories and several categories that are not present in novel or base categories, and also create corresponding anomalies. Results are presented in Tab. 1 and Tab. 2. NAS module achieves comparable performance across various potential categories, and, in some metrics, using only half of novel categories even achieves optimal performance. In all cases, the performance is superior to not using NAS module. This extensively validates the effectiveness of NAS module in open-world scenes.

3. t-SNE visualization of features

We show the t-SNE visualization of features in Fig. 1, we can see that vanilla visual features of CLIP reveal clear semantic identification but do not effectively distinguish different categories on VAD task. After specific optimization, these modified features used for the class-agnostic detection have more distinguishable boundaries between normal and abnormal distributions. Then, anomaly features used for the class-specific categorization exhibit fine-grained distinguishable boundaries across different categories and also surround the corresponding textual category embeddings. These t-SNE results clearly demonstrate that our model not only has good feature discrimination for the class-agnostic detection, but also has more refined feature recognition for the class-specific categorization.

4. Visualization of per-class results

To better verify the contribution of each module in our model, we present the detection and categorization results per class. For clarity, we arrange base categories at the front and novel categories at the back in each sub-figure of Fig. 2. It can be found that with the gradual introduction of TA, SKI, and NAS modules, the detection performance of almost every category is steadily improving, demonstrating that these specially devised modules are informative for the class-agnostic detection. For the class-specific categorization, two fine-tuning schemes yield different classification results. The fine-tuning without base anomalies tends to identify novel anomaly categories, thereby degrading the categorization performance of base anomaly categories. Therefore, we choose a trade-off scheme, i.e., fine-tuning our model with pseudo novel anomalies as well as base anomalies.

5. Qualitative novel anomaly generation results

We present in detail how to leverage off-the-shelf large models to generate pseudo novel anomalies in Fig. 3. As we can see, given the template containing anomaly categories, LLMs can generate detailed and authentic descriptions. Based on these descriptions, AIGC models create several plausible images. For example, in the third column, the image shows airbags that deployed after car collision, which appears in the corresponding description. Besides, it not hard to see that there is a gap in detailed information between these generated images and real images, but these

*Corresponding Authors

	AUC(%)	AUC_b(%)	AUC_n(%)	ACC(%)	ACC_b(%)	ACC_n(%)
50% Novel	86.13	93.80	87.98	40.02	45.09	37.07
50% Novel	**86.45**	**93.92**	**88.20**	38.57	43.17	35.92
50% Novel+Others	85.93	93.60	87.98	39.21	43.05	**37.14**
100%Novel	86.40	93.80	**88.20**	**41.43**	**49.02**	37.08

Table 1. Ablations studies on UCF-Crime with different potential categories for NAS module.

	AUC(%)	AUC_b(%)	AUC_n(%)	ACC(%)	ACC_b(%)	ACC_n(%)
50% Novel	**68.39**	58.78	**77.36**	**66.67**	**93.12**	30.36
50% Novel	65.49	61.29	73.50	60.26	91.98	16.75
50% Novel+Others	66.14	**61.48**	71.12	65.56	90.45	**31.41**
100% Novel	66.53	57.10	76.03	64.68	89.31	30.90

Table 2. Ablations studies on XD-Violence with different potential categories for NAS module.

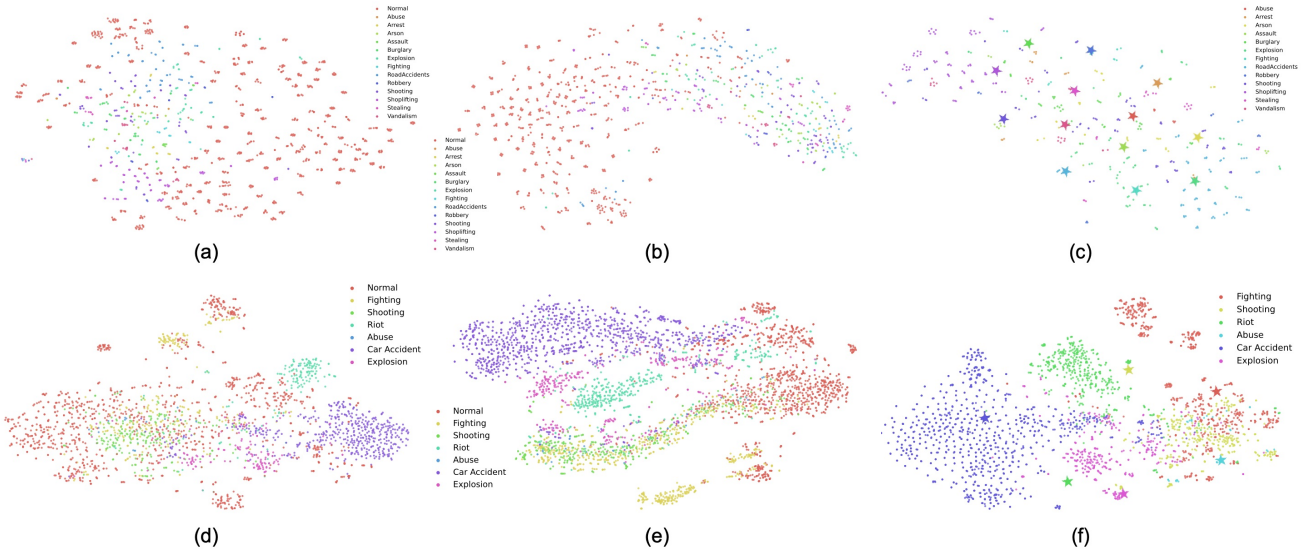

Figure 1. t-SNE visualizations on UCF-Crime (top) and XD-Violence (bottom). Left: Raw visual features of CLIP; Middle: Visual features used for the class-agnostic detection. Right: Visual features used for the class-specific categorization (star icons indicate textual category embeddings). Best view in color.

generated images still contain adequate semantics, which can assist task-specific model in improving its ability to perceive unseen categories.

References

[1] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 1

[2] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1

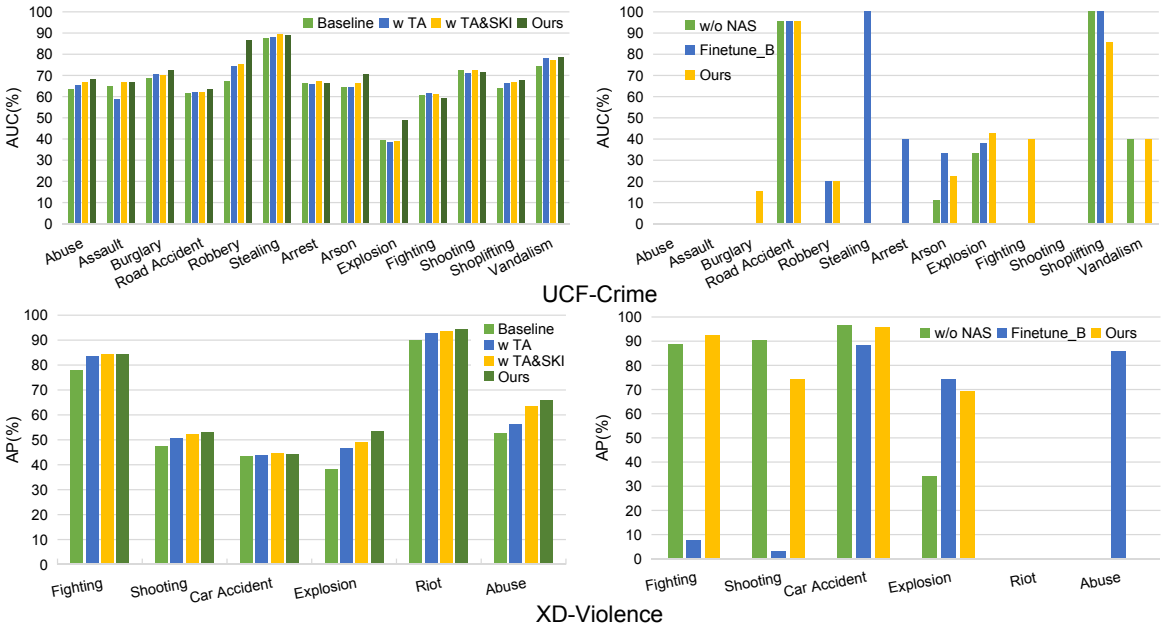

Figure 2. Contribution of each module to per-class open-vocabulary anomaly detection on UCF-Crime and XD-Violence. Left: Per-class detection results; Right: Per-class categorization results.

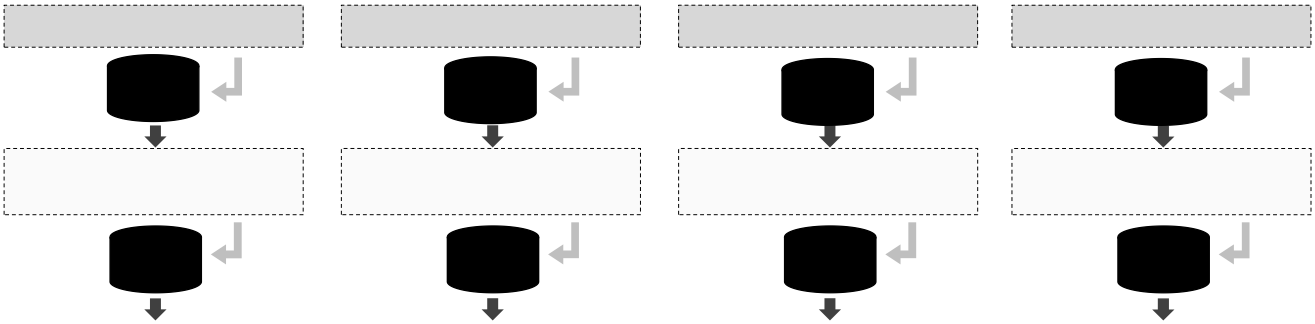

Figure 3. Qualitative results of novel anomaly generation process.