

# Supplementary Materials: Video Anomaly Detection via Progressive Learning of Multiple Proxy Tasks

Anonymous Authors

In this supplement, we provide the following:

- Exploration of feature ablation experiments conducted across different phases of progressive learning.
- Deliberation on the trade-off between training time cost and performance enhancement resulting from the progressive learning.
- Results from experiments encompassing a broader array of proxy tasks.
- Evaluation of the impact of diverse modal inputs (such as RGB images and optical flow) on progressive learning.
- Qualitative comparison with the SOTA methods and visualization.

## 0.1 Ablation Experiments on Progressive Learning

To validate the necessity of each phase in progressive learning, we conduct ablation experiments on the ShanghaiTech and Campus datasets. Meanwhile, to demonstrate the necessity of progressive learning, we supplement the results of simultaneous learning.

**Table 1: Ablation experiments on the multi-task learning. We report the AUC (%) scores on ShanghaiTech and Campus datasets. 'FP', 'SR', and 'Vir' stand for the proxy tasks of frame prediction, proposed semantic reconstruction, and virtual data-based classification, respectively. In addition, 'Prog' and 'Simu' represent the two training strategies of progressive and simultaneous learning, respectively.**

ID	FP	SR	Vir	Prog	Simu	AUC	
						ShTech	Campus
1	✓	-	-	-	-	73.1	57.9
2	-	✓	-	-	-	76.1	65.4
3	-	-	✓	-	-	68.8	58.1
4	✓	✓	-	✓	-	77.2	65.4
5	✓	-	✓	✓	-	75.1	59.9
6	-	✓	✓	✓	-	78.8	65.2
7	✓	✓	✓	✓	-	79.0	67.2
8	✓	✓	-	-	✓	79.2	66.2
9	✓	-	✓	-	✓	72.0	57.0
10	-	✓	✓	-	✓	79.7	66.9
11	✓	✓	✓	-	✓	86.2	69.4

As shown in table 1, each phase of the proposed framework is indispensable. The frame prediction task in the perception phase learns low-level pixel features to lay the foundation for learning in subsequent phases. Either progressive learning of subsequent semantic proxy tasks or simultaneous learning can improve the performance of the model (ID 2, 4, 8). The semantic reconstruction

task in the comprehension phase learns semantic features, as evidenced by the significant performance gains of the model on the Campus dataset (ID 1, 4, 8). After completing the first two phases of learning, the model is trained using virtual data in the inference phase helps to learn general features to improve model detection (ID 11). In contrast, learning three tasks simultaneously leads to model convergence to the sub-optimal point, and the model performance is rather inferior to learning two tasks simultaneously (ID 4, 7). Progressive learning effectively avoids this problem, and the performance of the model continues to improve over the three phases (ID 8, 11).

It is important to note that performing the frame prediction task first and training with virtual data immediately after can lead to a degradation in model performance (ID 9). This is due to the fact that the model performs difficult tasks early in training leading to challenges that are difficult for the model to solve. This also proves the necessity of the three stages and the training sequence "Perception - Comprehension - Inference". Performing the semantic reconstruction task first and then training with virtual data does not present a similar problem (ID 10).

## 0.2 Trade-off between Time Cost and Performance Gain

In order to avoid convergence of the model to a sub-optimal point, we train with different proxy tasks at different phases. For continuous performance improvement, we propose progressive learning to provide different but successive optimization goals. The process of progressive learning in fact transforms simultaneous learning into multi-phase learning. Multi-phase learning inevitably induces an increase in the cost of training time, and we explore the trade-off between training cost and performance gain.

Three tasks of frame prediction, semantic reconstruction, and virtual data-based classification are used as examples to perform progressive learning. As shown in Table 1, the model obtained from progressive learning training achieves AUC scores of 86.2% on the ShanghaiTech dataset. While, the model that learns three tasks simultaneously achieves AUC scores of 79.0%. Progressive learning brings performance gains of up to 7.2%.

With an NVIDIA RTX 3090 GPU, it takes 80 epochs for the model to learn three tasks simultaneously for training. With the progressive learning, the frame prediction (Perception Phase), semantic reconstruction (Comprehension Phase) and virtual data-based classification (Inference Phase) tasks require 30, 40 and 20 epochs, respectively. Hence, the slight increase in training time is justified by the significant performance improvements facilitated by progressive learning. Moreover, the cumulative training duration across the three phases of progressive learning does not simply triple compared to simultaneous learning, indicating that the objectives achieved in the initial phases contribute to the efficiency of subsequent phases.

**Table 2: AUC (%) performance of models trained with different combinations of tasks using different learning sequences and models trained simultaneously with the same weights for all tasks on the ShanghaiTech and Campus datasets. The results indicate that our progressive learning approach achieves the maximum performance improvement when learning multiple auxiliary tasks, and the performance gains are not limited to specific tasks.**

Dataset	Progressive						Simultaneous
	Learning Order			AUC			AUC
	Phase 1	Phase 2	Phase 3	Phase 1	Phase 2	Phase 3	
ShanghaiTech	Pre	→SR	→ Virtual	73.1	79.2	84.8	79.0
	Pre	→Virtual	→SR	73.1	72.0	75.5	74.2
	SR	→Pre	→Virtual	76.1	75.5	76.8	76.3
	CLS	→Virtual	→SR	65.4	64.2	73.2	70.1
	-	→Virtual	→SR	-	68.8	76.2	72.1
	Rec	→Jigsaw	→VL	71.9	81.2	82.3	80.2
	Pre	→Optical Flow	→Virtual	73.1	76.8	78.8	77.2
	Pre	→Optical Flow→SR	→Virtual	73.1	80.2	85.3	79.2
	Rec	→Completion	→VL	71.9	79.8	81.9	79.5
	Rec	→Completion	→Pseudo	71.9	79.8	81.1	80.0
	Pre → Rec	-	→ Pseudo	76.0	76.0	79.3	79.0
	Pre → Rec	→ SR	-	76.0	85.0	85.0	82.3
	Pre → Rec	→ Jig	→ Virtual	76.0	85.9	88.6	79.2
	Pre → Rec	→ Jig	→ Virtual→ VL	76.0	85.9	88.7	79.5
	Pre → Rec	→ Jig → Completion	→ Virtual	76.0	86.1	88.8	78.1
	Pre → Rec → CLS	→ SR → Jig → Completion	→ Virtual	76.8	86.4	88.8	77.9
Campus	Pre	→SR	→ Virtual	57.9	66.2	69.4	67.2
	Pre	→Virtual	→SR	57.9	57.0	68.1	67.2
	SR	→Pre	→Virtual	65.4	64.8	66.1	65.9
	CLS	→Virtual	→SR	54.2	53.9	62.8	62.0
	-	→Virtual	→SR	-	58.1	64.3	64.2
	Rec	→Jigsaw	→VL	55.1	71.2	72.1	68.0
	Pre	→Optical Flow	→Virtual	57.9	58.8	59.4	59.4
	Pre	→Optical Flow→SR	→Virtual	57.9	66.5	69.6	67.7
	Rec	→Completion	→VL	55.1	67.9	70.6	67.2
	Rec	→Completion	→Pseudo	55.1	67.9	71.1	67.8
	Pre → Rec	-	→ Pseudo	58.2	58.2	60.6	61.3
	Pre → Rec	→ SR	-	58.2	69.9	69.9	66.1
	Pre → Rec	→ Jig	→ Virtual	58.2	71.2	73.3	70.2
	Pre → Rec	→ Jig	→ Virtual→ VL	58.2	71.2	73.5	70.0
	Pre → Rec	→ Jig → Completion	→ Virtual	58.2	72.3	75.1	66.9
	Pre → Rec → CLS	→ SR → Jig → Completion	→ Virtual	59.1	74.5	75.8	66.8

### 0.3 More Proxy Task Combinations and Training Sequences

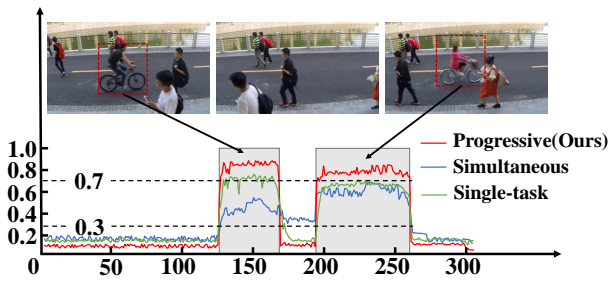
In the ablation experiments in the main text, we report experimental results for different proxy task combinations and training sequences. In this Supplementary Material, we report the performance of more proxy task combinations and training sequences on the ShanghaiTech and Campus datasets.

In keeping with the main text, we select the widely used proxy task. For the visual proxy task, we select the reconstruction [5] and classification [5] tasks. For the semantic proxy task, we select the event completion [9] and puzzle task [6]. For the open-set proxy task, we select the background-agnostic [1] of synthesizing pseudo anomalies and the Vision-Language task [8] of generating novel anomalies using pre-trained multi-model model.

As the results in Table 2 show, proxy tasks such as reconstruction, classification, event completion, and pseudo-data are all consistently improved in model performance when divided and ordered with the guidelines of progressive learning. More comprehensive experiments demonstrate that the performance gains from progressive learning are not limited to specific tasks.

### 0.4 Ablation Experiment on Diverse Modal Inputs

In both the main text and the above experiments, the input to our model is RGB images. In fact, there are many SOTA methods [2–4] that use optical flow images as additional input and design proxy tasks. Therefore, we further explore the enhancement utility of optical flow images in progressive learning.



**Figure 1: Frame-level scores example for test video from ShanghaiTech dataset.** The vertical and horizontal axes represent anomaly scores and video frames respectively. We show the sampled video frames, the anomalous portion of the ground truth (gray areas are anomalies), and the anomaly scores achieved by the anomaly detection model. The red, green, and blue curves represent the results of progressive learning of multiple proxy tasks (Ours), learning multiple proxy tasks simultaneously, and single-task design [7], respectively. Best viewed in color.

Previous studies typically reconstruct optical flow images to learn motion features (which are semantic features). Therefore, we are consistent with previous studies that perform optical flow-based reconstruction in the comprehension phase. As shown in Table 2, the optical flow images can be unified in our proposed progressive learning to continuously improve the performance of the model. Nonetheless, it’s important to note that optical flow information primarily encapsulates semantic details regarding motion patterns rather than scene information. Therefore, its augmentative effect on the Campus dataset isn’t as pronounced as observed with other semantic proxy tasks.

## 0.5 Qualitative Experiments and Visualization

Figure 1 shows the anomaly score curves obtained by progressive learning on the ShanghaiTech dataset. To illustrate the effectiveness of proposed method, we compare our method with simultaneous learning and the SOTA single-task method (ROADMAP [7]).

As shown in the figure, the model trained with progressive learning rises faster when there are anomalies and has a higher anomaly score. When the video is normal, the score is even lower and smoother. The proxy tasks performed by both progressive and simultaneous learning are frame prediction, semantic reconstruction, and virtual data-based classification.

## REFERENCES

- [1] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. 2022. A Background-Agnostic Framework With Adversarial Training for Abnormal Event Detection in Video. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 9 (2022), 4505–4523.
- [2] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. 2019. Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video. In *CVPR*. 7842–7851.
- [3] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 2018. Future Frame Prediction for Anomaly Detection - A New Baseline. In *CVPR*. 6536–6545.
- [4] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guoqing Li. 2021. A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow

- Reconstruction and Flow-Guided Frame Prediction. In *ICCV*. 13568–13577.
- [5] Chenrui Shi, Che Sun, Yuwei Wu, and Yunde Jia. 2023. Video Anomaly Detection via Sequentially Learning Multiple Pretext Tasks. In *ICCV*. 10296–10306.
- [6] Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, and Di Huang. 2022. Video Anomaly Detection by Solving Decoupled Spatio-Temporal Jigsaw Puzzles. In *ECCV*. 494–511.
- [7] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. 2022. Robust Unsupervised Video Anomaly Detection by Multipath Frame Prediction. *IEEE Trans. Neural Networks Learn. Syst.* 33, 6 (2022), 2301–2312.
- [8] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. 2023. Open-Vocabulary Video Anomaly Detection. *CoRR* abs/2311.07042 (2023).
- [9] Guang Yu, Siqu Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. 2020. Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events. In *ACM Multimedia*. 583–591.