VideoPatchCore: An Effective Method to Memorize Normality for Video Anomaly Detection

Sunghyun Ahn

Department of Computer Science, Yonsei University

Data Engineering Laboratory

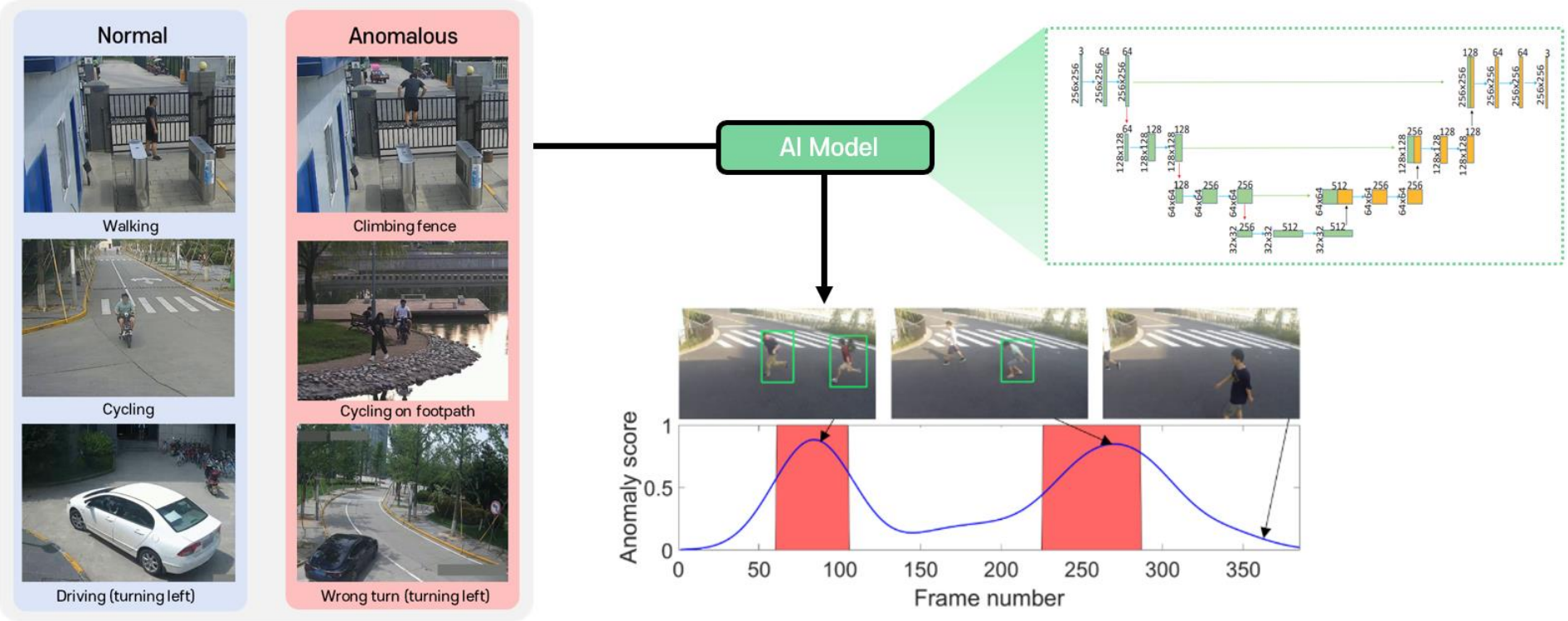
M.S. Thesis Presentation

00 Outline

- Background
- Introduction
- Method
- Experiments
- Conclusion
- Appendix

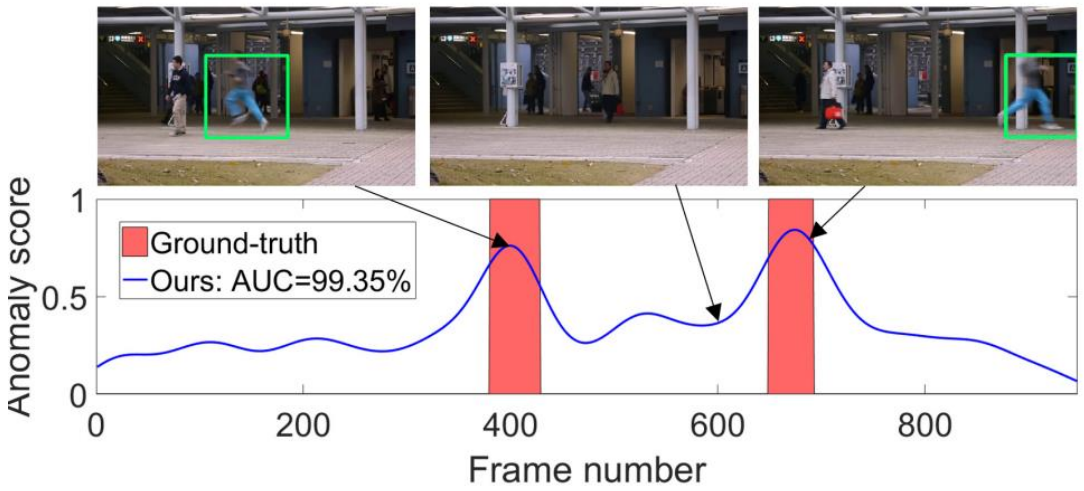
01 Background Video Anomaly Detection

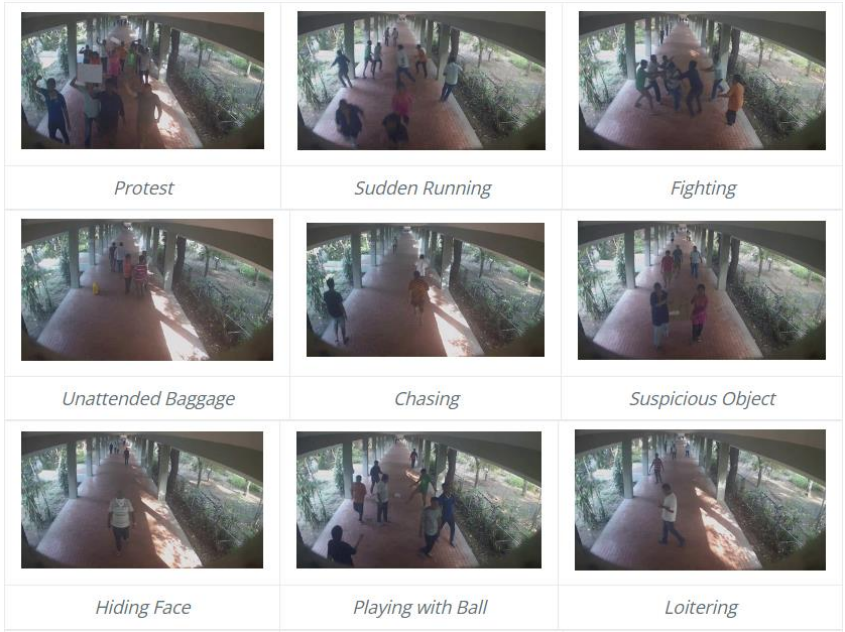
- **Video Anomaly Detection** aims to determine whether abnormal events occur within video streams
- Abnormal events include the **appearance** or **action** of objects that are not suitable for the situation.
- The goal is to do **Binary Classification**.



01 Background Main Problem

- Class imbalance problem $|\{x_i|y_i = 0\}| \gg |\{x_i|y_i = 1\}|$
- Diverse anomaly
- **one-class classification** is utilized that learns exclusively from normal data and classifies anything not resembling the patterns of normal data as abnormal.





Rodrigues, Royston, et al. "Multi-timescale trajectory prediction for abnormal human activity detection." *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2020.

01 Background One Class Classification

- **Classification-based**: Learning cluster that group normal feature vectors within a specific range.

$$\mathcal{A}_\theta(x) = \mathrm{Diff}(\phi_\theta(x), c)$$

Key idea
- normal data: clustered in a 'normal' cluster
- abnormal data: away from the 'normal' cluster

ϕ_θ: kernel function
c: center of hypersphere

- **Distance-based**: Storing features of normal data in memory and using the Nearest Neighbor search for classification.

$$\mathcal{A}_\theta(x) = \min_{x_{tr} \in X_{tr}} \mathrm{Diff}\big(\phi_\theta(x), \phi_\theta(x_{tr})\big)$$

Key idea
- normal data: similar to training data in a feature space
- abnormal data: NOT similar to training data in a feature space

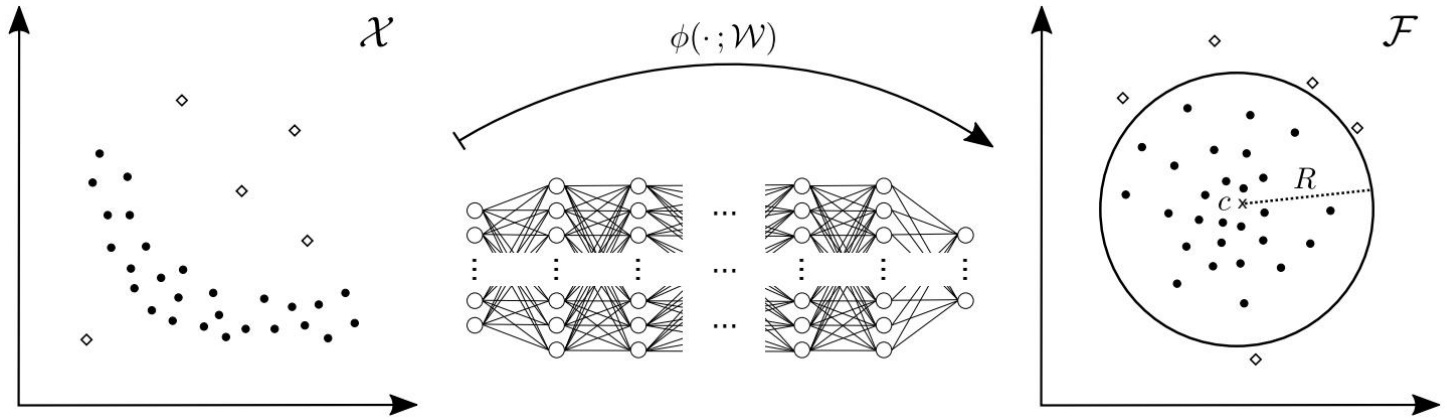
- **Reconstruction-based**: Learning to reconstruct normal samples using a generative model.

$$\mathcal{A}_\theta(x) = \mathrm{Diff}(x, \mathrm{Recon}_\theta(x))$$

Key idea
- normal data: reconstructed well
- abnormal data: NOT reconstructed well

01 Background Classification-based Method (Deep SVDD)

- Find the smallest **hypersphere** that surrounds the normal data in the feature space and use its boundary to detect anomalies.
- Learn the kernel function that maps normal features inside the hypersphere through deep learning.
- Anomaly Score is the degree to which the output feature is distant from the center.



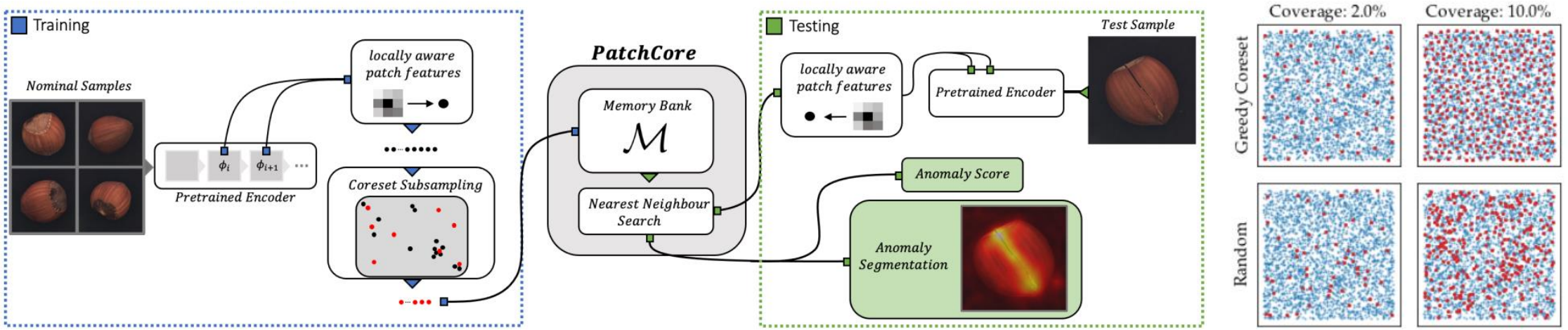
$$\min_{\mathcal{W}} \quad \frac{1}{n}\sum_{i=1}^{n}\|\phi(\boldsymbol{x}_i;\mathcal{W})-\boldsymbol{c}\|^2+\frac{\lambda}{2}\sum_{\ell=1}^{L}\|\boldsymbol{W}^\ell\|_F^2.$$

$$s(\boldsymbol{x}) = \|\phi(\boldsymbol{x};\mathcal{W}^*)-\boldsymbol{c}\|^2$$

Ruff, Lukas, et al. "Deep one-class classification." International conference on machine learning. PMLR, 2018.

01 Background Distance-based Method (PatchCore)

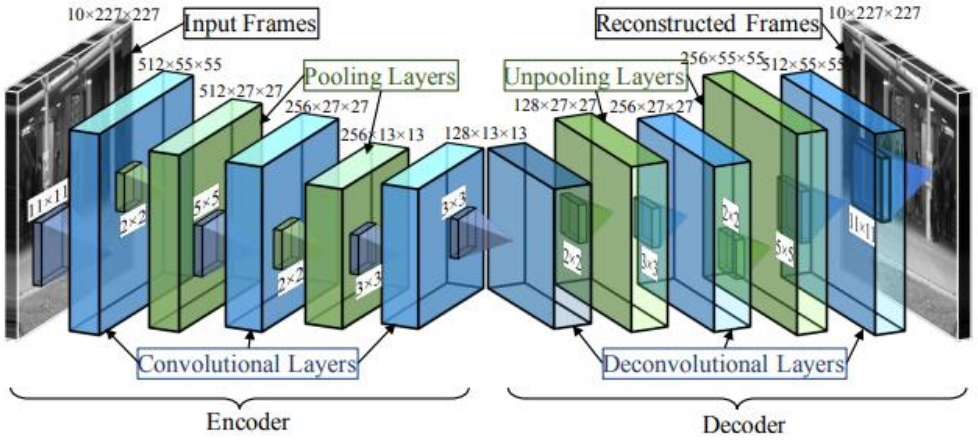
- Utilize a pre-trained model to **memorize the normal features** and apply this memory for anomaly detection.
- Storing all normal features in memory can lead to **hardware constraints**.
- Features stored in memory are subsampled using a **greedy coreset subsampling algorithm**, where a coreset is a small amount of data that well represents the existing dataset.



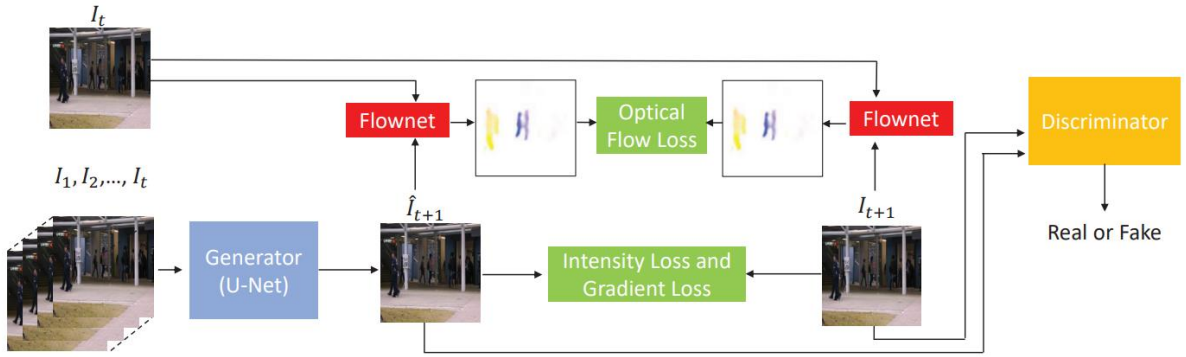
Roth, Karsten, et al. "Towards total recall in industrial anomaly detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

01 Background Reconstruction-based Method (Conv-AE, FFP)

- **Conv-AE** assumes that the autoencoder is trained to reconstruct only normal frames, meaning it will not be able to reconstruct anomalous frames.
- **FFP** assumes that if it is trained to predict only normal future frame, it will not be able to predict anomalous future frame.
- Deep learning models can generate anomalous frames due to their powerful generalization capability.



Hasan, Mahmudul, et al. "Learning temporal regularity in video sequences." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

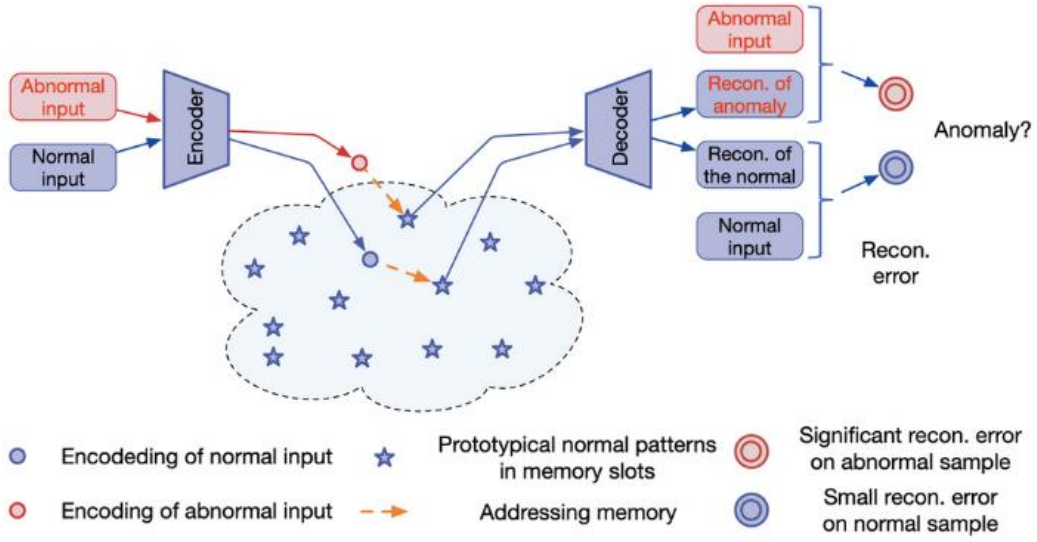


Liu, Wen, et al. "Future frame prediction for anomaly detection–a new baseline." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

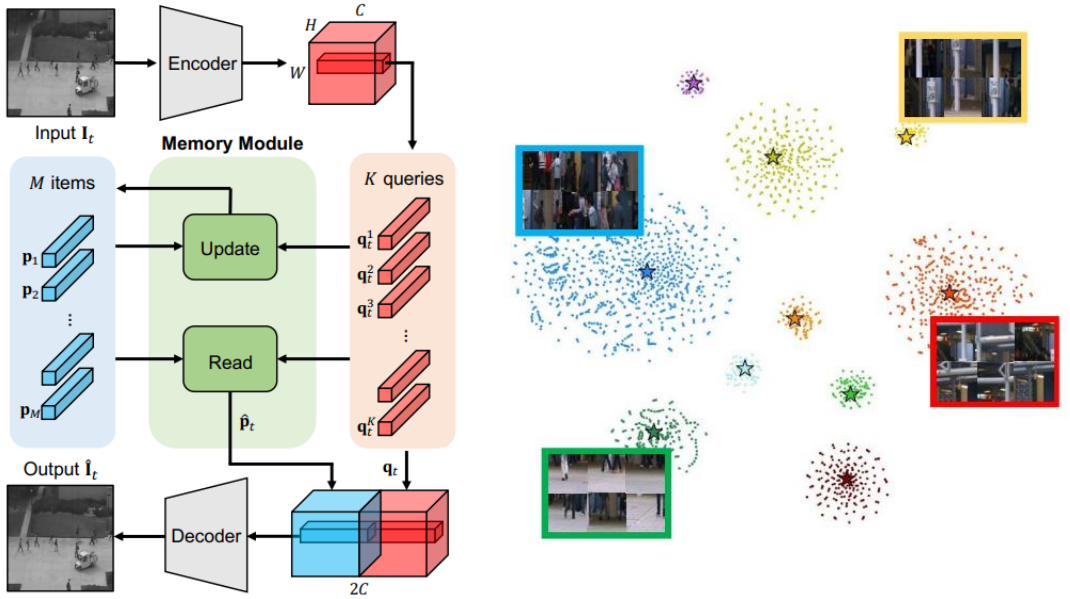
01 Background Reconstruction-based Method (MemAE, MNAD)

- **MemAE** stores the features of normal data in a memory module and **generate normal data using memory**.
- **MNAD** attempts anomaly detection using only M memory items to generate normal data.

  The model is trained to cluster normal features based on memory items to create independent items.



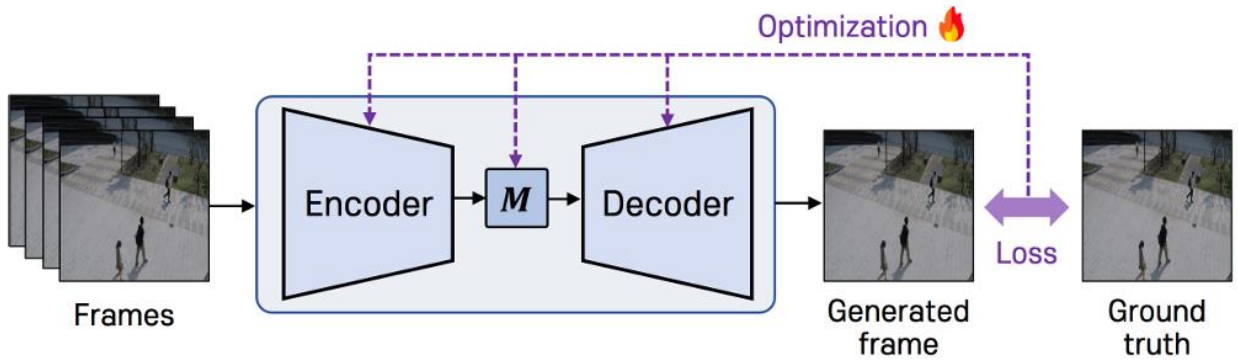
Gong, Dong, et al. "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection.“
Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.



Park, Hyunjong, Jongyoun Noh, and Bumsub Ham. "Learning memory-guided normality for anomaly detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

02 Introduction Problem Definition

- **Optimizing** the memory and encoder-decoder simultaneously is **challenging** due to different training objectives.
- **Developing** both the memory and model is **complex**, making it harder to approach the VAD task.
- **Memory size** can significantly **impact performance**: a smaller size may lower performance, while a larger size can increase latency.



	Methods	Ped2 [21]	Avenue [24]	Shanghai [26]
-	MPPCA [15]	69.3	-	-
	MPPC+SFA [15]	61.3	-	-
	MDT [28]	82.9	-	-
	AMDN [46]	90.8	-	-
	Unmasking [41]	82.2	80.6	-
	MT-FRCN [10]	92.2	-	-
	AMC [31]	96.2	86.9	-
Recon.	ConvAE [9]	85.0	80.0	60.9
	TSC [26]	91.0	80.6	67.9
	StackRNN [26]	92.2	81.7	68.0
	AbnormalGAN [33]	93.5	-	-
	MemAE w/o Mem. [8]	91.7	81.0	69.7
	MemAE w/ Mem. [8]	94.1	83.3	71.2
	Ours-R w/o Mem.	86.4	80.6	65.8
	Ours-R w/ Mem.	90.2	82.8	69.8
Pred.	Frame-Pred [22]	95.4	85.1	**72.8**
	Ours-P w/o Mem.	94.3	84.5	66.8
	Ours-P w/ Mem.	**97.0**	**88.5**	70.5

Park, Hyunjong, Jongyoun Noh, and Bumsub Ham. "Learning memory-guided normality for anomaly detection." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

02 Introduction Key Idea

❖ **Objectives**

➢ Explore effective memory utilization method for Video Anomaly Detection.

❖ **Success in the Image Domain: PatchCore (CVPR 2022)**

① Optimize memory only and attempt anomaly detection through distance comparison between memory and features.

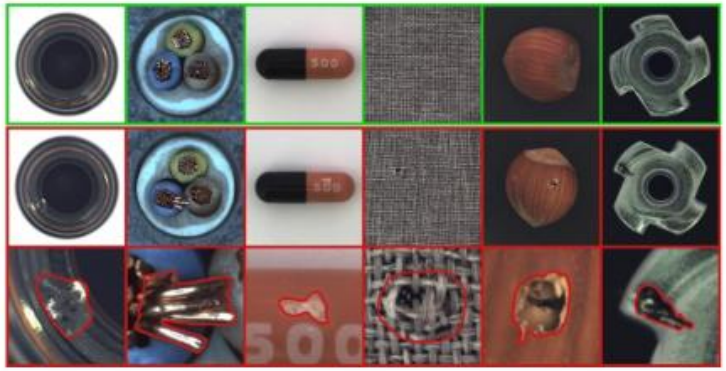
② Ensure robust performance against memory size using Coreset Subsampling.

❖ **Extending PatchCore to the Video Domain**

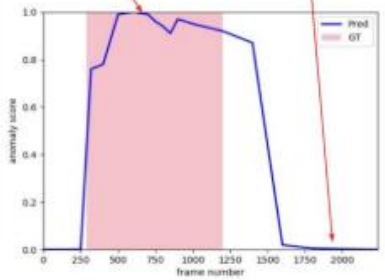
➢ Development of VideoPatchCore enables effective memory utilization for anomaly detection in the video domain.

02 Introduction Extending to the video domain

- Video is composed of a continuous sequence of frames.
- Anomaly is not limited to a single object but can be determined by the interactions between multiple objects.
- There are anomalous events that can only be determined by observing the scene (i.e., wrong direction).



VS

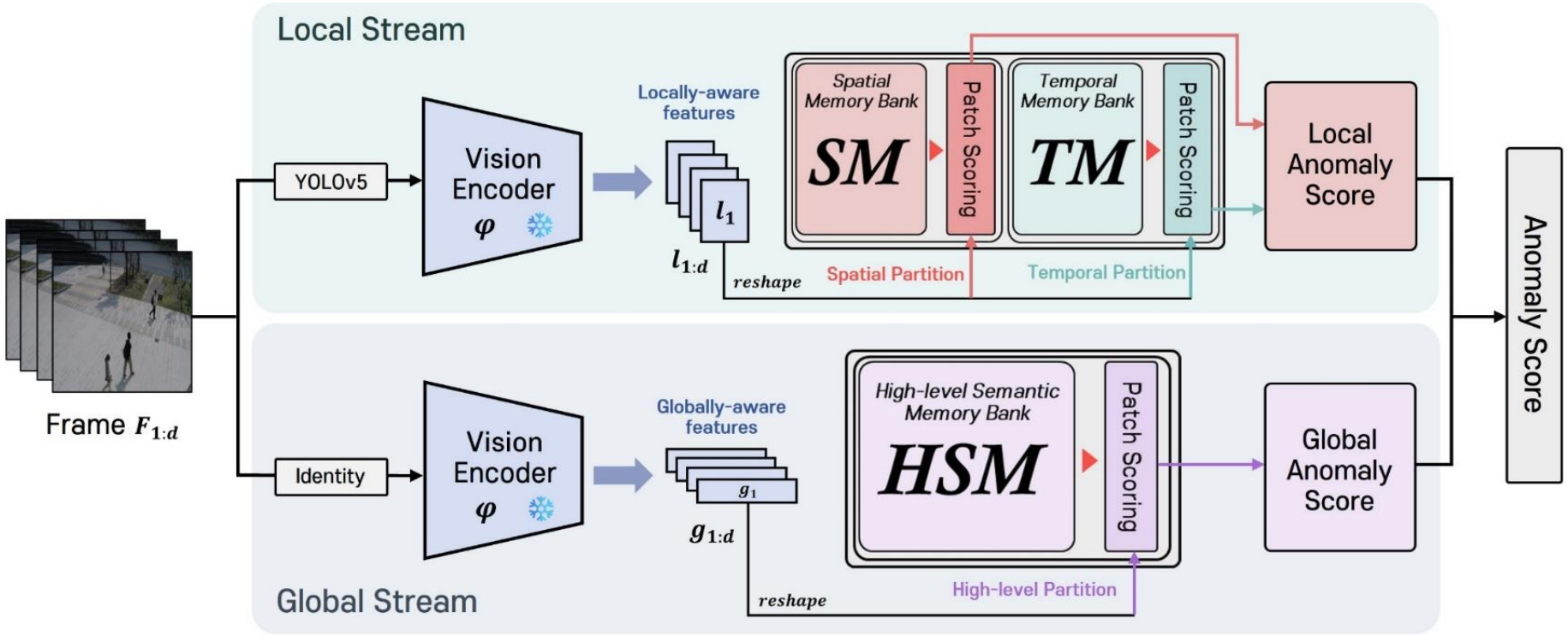


Fighting

- **Utilize temporal information (i.e., motion) to represent the time-varying changes of objects.**
- **Perform local anomaly detection at the object level and conduct global anomaly detection at the frame level.**

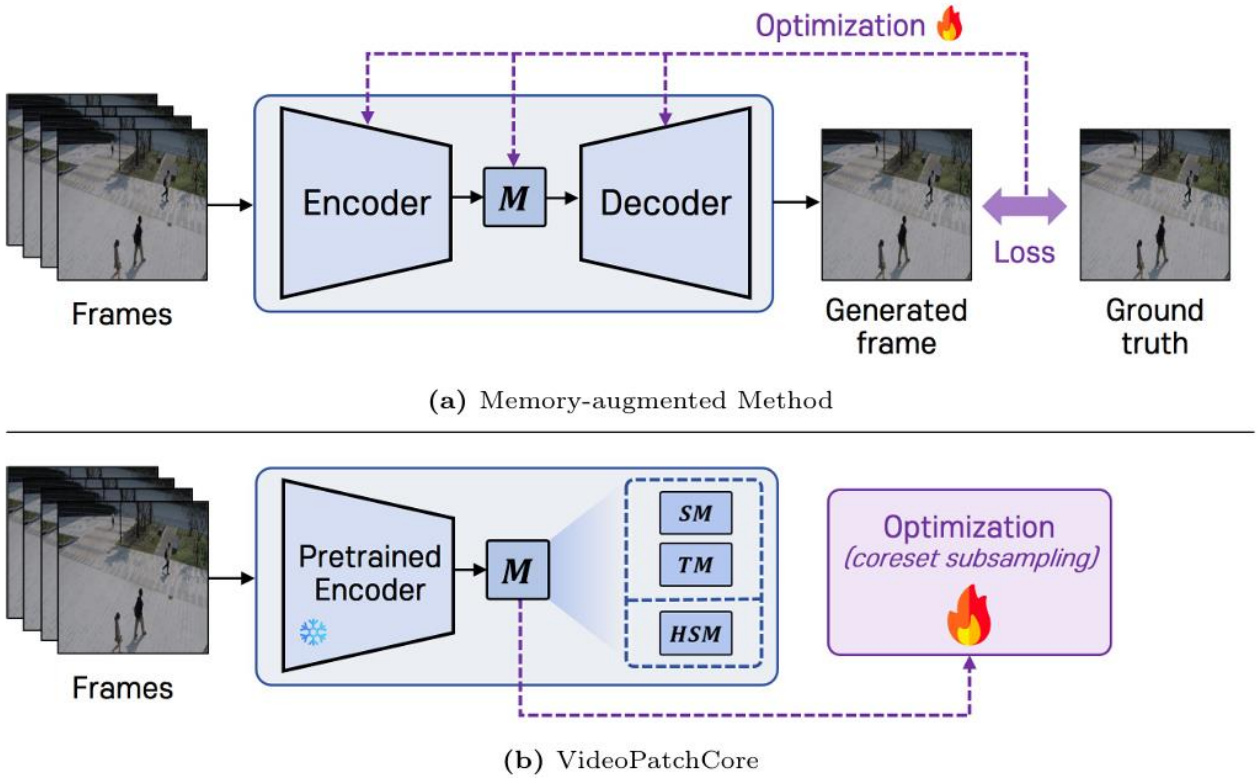
02 Introduction Overview

- An effective approach for memorizing normal features for VAD.
- **Local stream** is object-based, detecting **anomalous appearances and actions of individual objects.**
- **Global stream** is frame-based, detecting **anomalies related to multiple objects or scenes.**



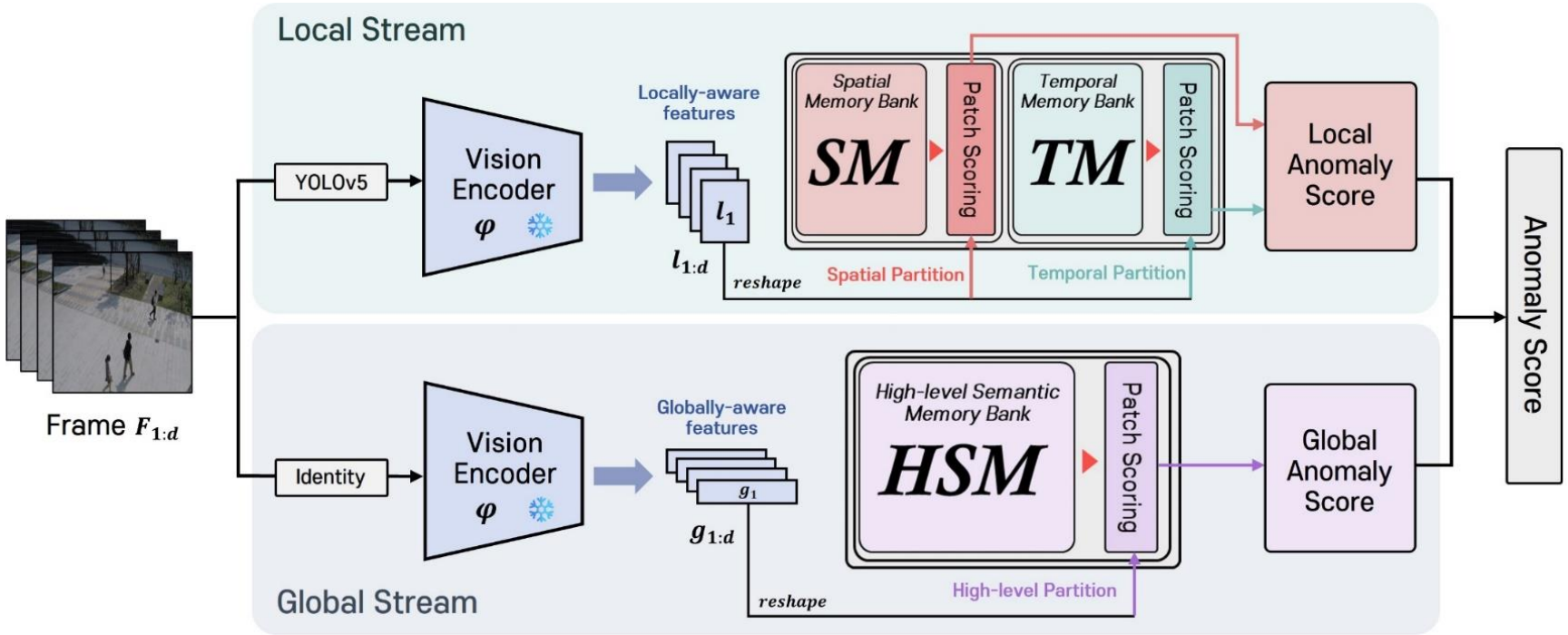
02 Introduction Argument

- **The existing method (Memory-augmented Method)** that optimizes both the model and memory has three issues: increased optimization difficulty, complexity of implementation, and performance variability depending on the memory size.
  The **proposed method (VideoPatchCore),** which optimizes only the memory, can address all these issues.



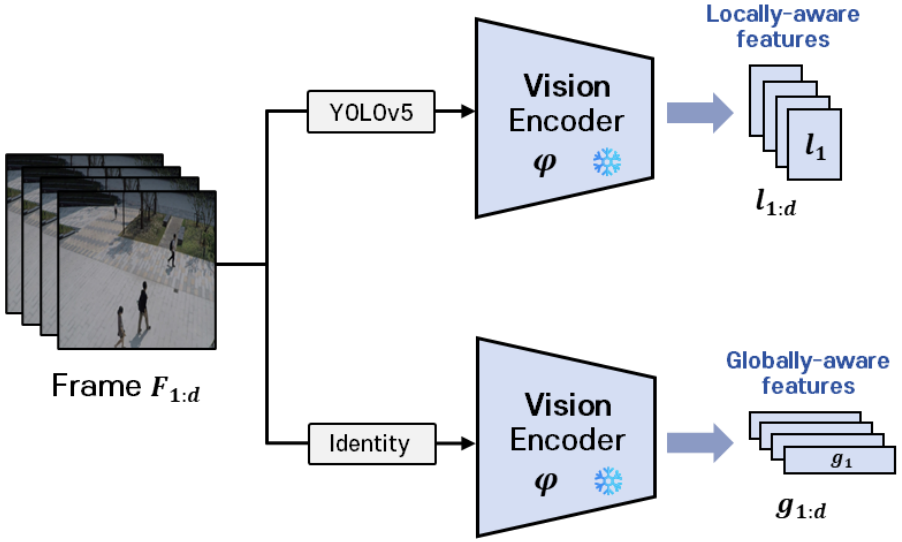
03 Method Two stages

- There are **two stages**: Memorization and Inference.
- In the **memorization stage**, features are stored in memory, which is then optimized using greedy coreset subsampling from PatchCore.
- In the **inference stage**, anomaly scores are derived by calculating distances between memory and features.



03 Method Vision Encoder

- CNN-based CLIP model is adopted as the encoder, utilizing layer 2 and layer 3 (φ_2, φ_3) similar to PatchCore.
- **Locally-aware features** $l_{1:d}$ represent both fine- and coarse-grained information of the objects.
- **Globally-aware features** $g_{1:d}$ represent global information of the frames.



$$l_i = \langle f_{ap}(\varphi_2(O_i)), f_{ap}(\varphi_3(O_i)) \rangle \quad (1)$$

where f_{ap} represents the average pooling and $\langle \cdot, \cdot \rangle$ denotes tensor concatenation. Subsequently, $l_{1:d}$ are reshaped into $lf \in \mathbb{R}^{n \times c \times d \times h \times w}$.

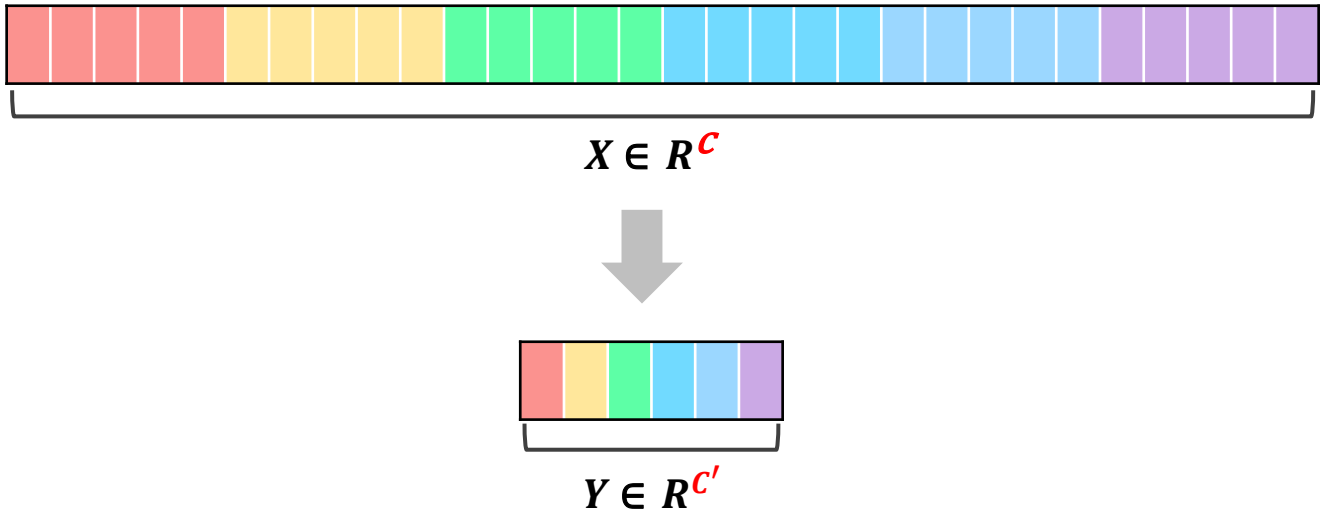
$$g_i = \langle f_{ap}(\varphi_2(F_i)), f_{ap}(\varphi_3(F_i)) \rangle \quad (2)$$

$$g_i = f_{gap}(g_i) + f_{gmp}(g_i) \quad (3)$$

where f_{gap} and f_{gmp} denote global average pooling and global max pooling, respectively. Subsequently, $g_{1:d}$ are reshaped into $gf \in \mathbb{R}^{c \times d \times 1 \times 1}$.

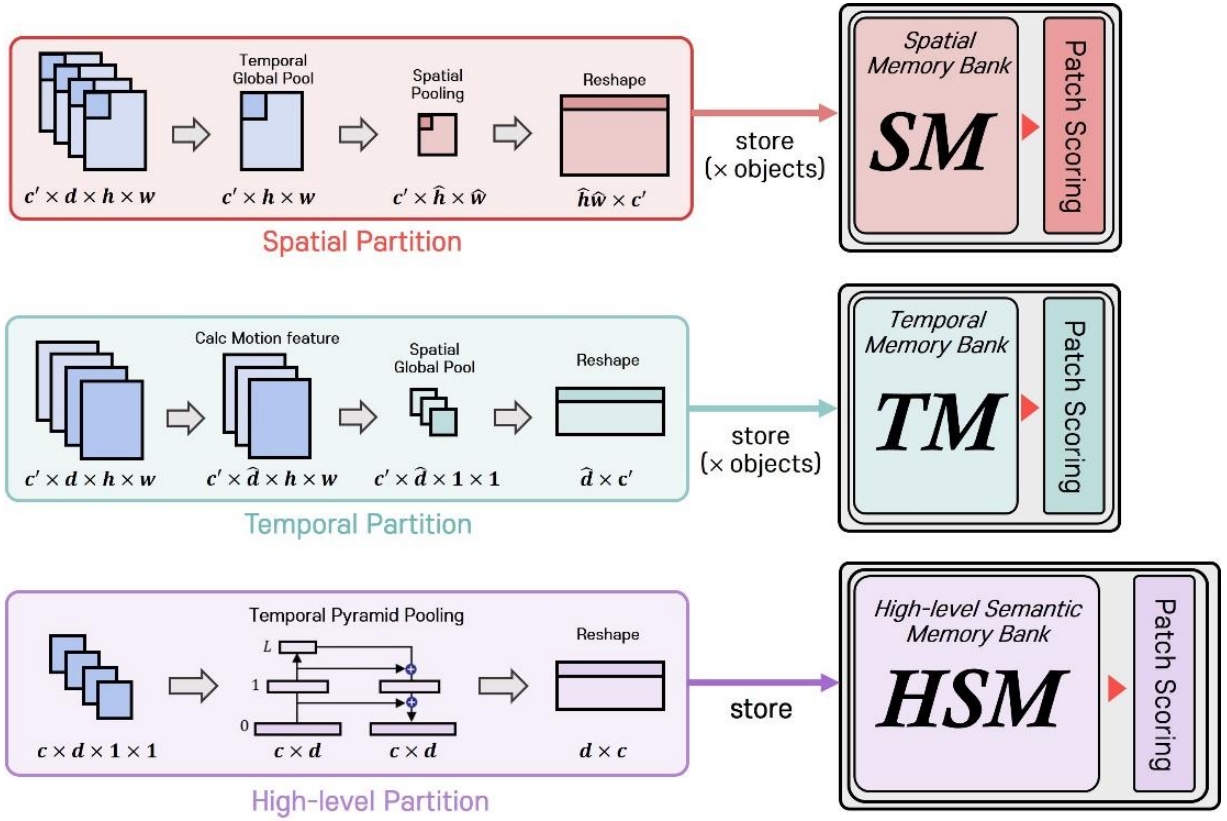
03 Method Split Pooling

- The dimensionality of $lf \in \mathbb{R}^{n \times c \times d \times h \times w}$ is impractically large for video processing.
- **Split pooling** reduce the number of channels, transforming lf into $\mathbb{R}^{n \times c' \times d \times h \times w}$ $(c \gg c')$
- Divide lf into c' groups and compresses by averaging the channels of each group.



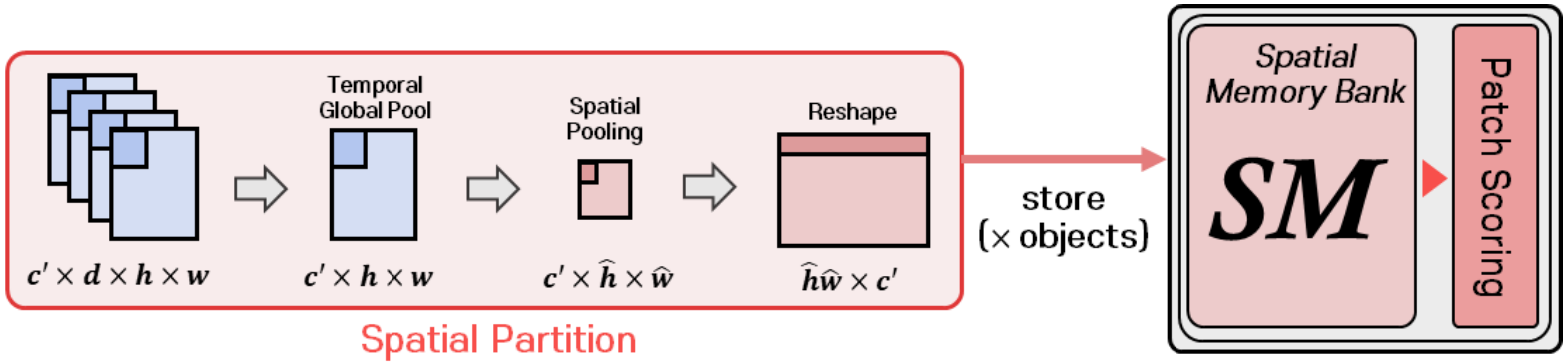
03 Method Patch Partition

- **Spatial partition** focuses on appearance information, generating patches while disregarding temporal information.
- **Temporal partition** emphasizes motion information, generating patches while ignoring spatial information.
- **High-level partition** utilizes extensive spatiotemporal features for extracting the global context across frames.



03 Method Spatial Partition

- Appearance information is crucial for evaluating anomalies in objects and can be derived from spatial features.
- Temporal global pooling preserves spatial information while ignoring temporal aspects.
- Average pooling is applied to consider various regions of the object.

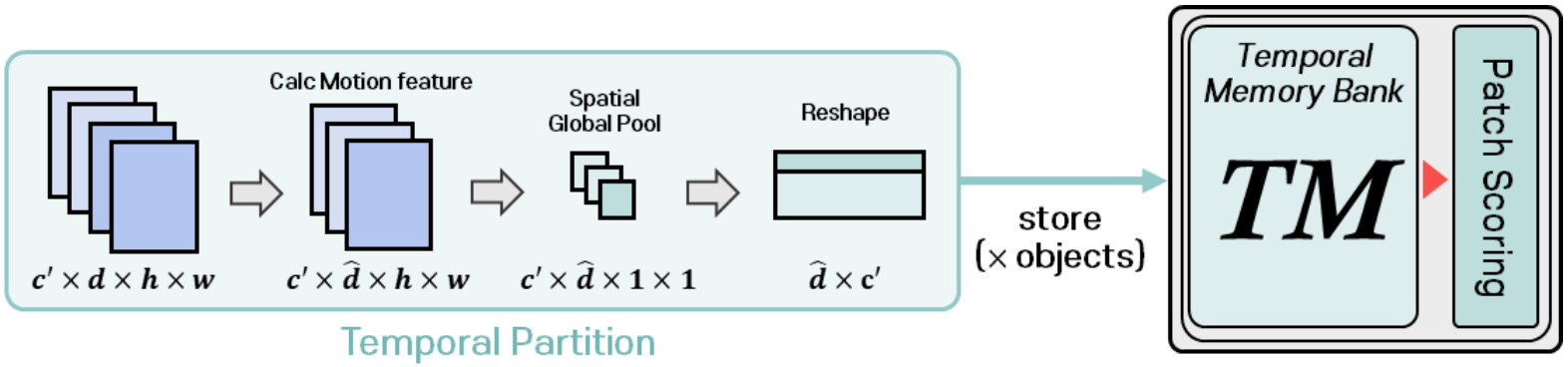


$$SpatialPatches = f_{ap}\left(f_{tgp}(lf)\right) \tag{4}$$

where f_{tgp} denotes the temporal global pooling. Finally, the results are reshaped into $\mathbb{R}^{(n\cdot\hat{h}\cdot\hat{w})\times c'}$, where $h \geq \hat{h}$ and $w \geq \hat{w}$.

03 Method Temporal Partition

- Motion information in objects represents changes over time, making it crucial for VAD.
- Utilize the adjacent temporal information within lf to generate motion features mf.
- Feature differences are computed and the representative motion values are determined through global pooling.



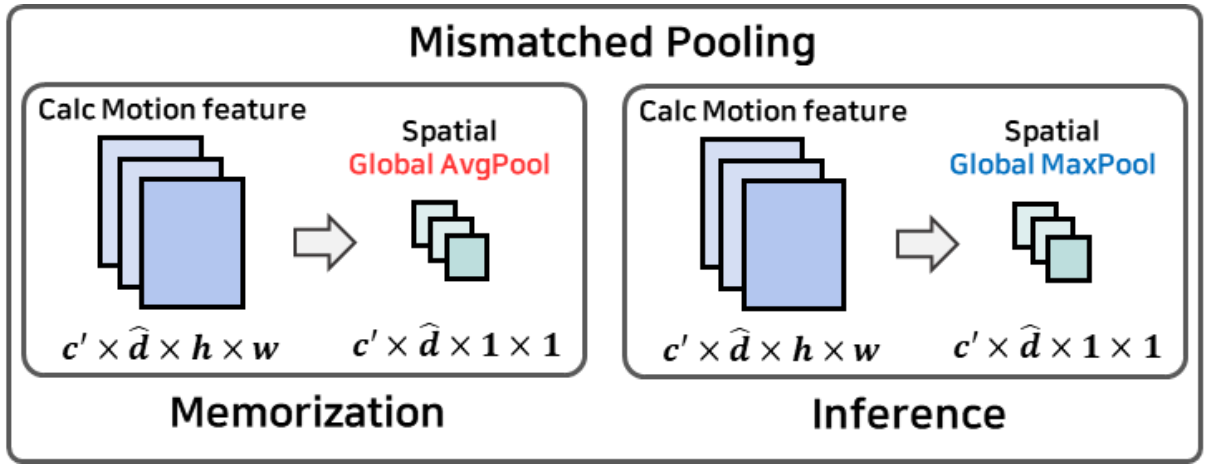
$$mf_{(t)} = \left| lf_{(t+1)} - lf_{(t)} \right| \quad (5)$$

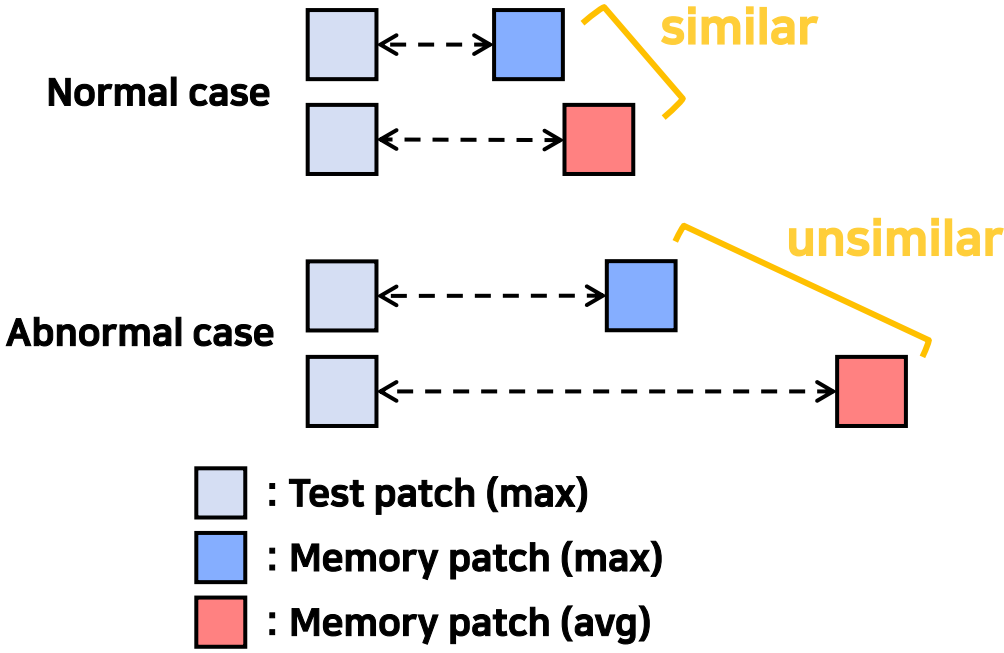
$$TemporalPatches = f_{gap}(mf) \quad (6)$$

$mf_{(t)} \in \mathbb{R}^{n \times c' \times h \times w}$ represents the difference between the t^{th} and $t+1^{th}$ time within lf, and it belongs to $mf \in \mathbb{R}^{n \times c' \times \hat{d} \times h \times w}$, where $\hat{d} = d - 1$. Finally, the results are reshaped into $\mathbb{R}^{(n \cdot \hat{d}) \times c'}$.

03 Method Mismatched Pooling

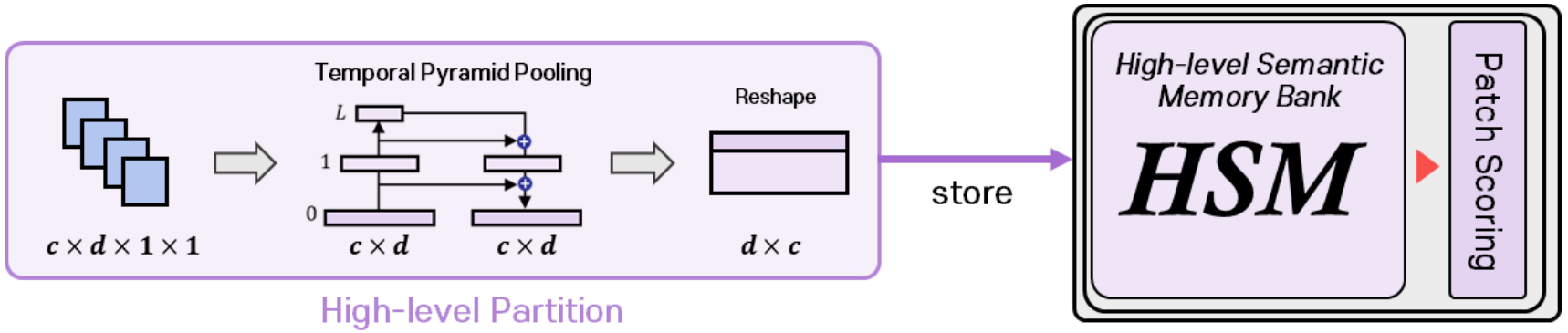
- **Mismatched Pool** performs GAP during the memorization and GMP during the inference stage.
- Normal objects are more static than abnormal ones,
  so the average and maximum of normal motion features are expected to be similar.
  In contrast, abnormal motion features are expected to show significant differences between these values.
- Prevent the retrieval of nearby patch from memory when abnormal patch is input.





03 Method High-level Partition

- In frames, Global context processes the relationship between the objects and the scene, and considers interactions between different objects.
- **Temporal pyramid pooling** is utilized to obtain high-level temporal information.
- Secure multi-scale temporal information, addressing the limitation of only using adjacent temporal information.



High-level Partition

$$HighlevelPatches = \sum_{l=0}^{L} f_{mp}^{l}(gf) \quad (7)$$

Temporal pyramid pooling is implemented using f_{mp}^{l}. In this case, f_{mp}^{l} represents applying the max pooling operation l times. Finally, the results are reshaped into $\mathbb{R}^{d \times c}$.

03 Method Anomaly Scoring

- Using the nearest neighbor method, the closest memory item to the patch is found, and the maximum patch score s^* is computed.
- The final anomaly score is determined by calculating the weighted sum of the scores computed for each memory bank.

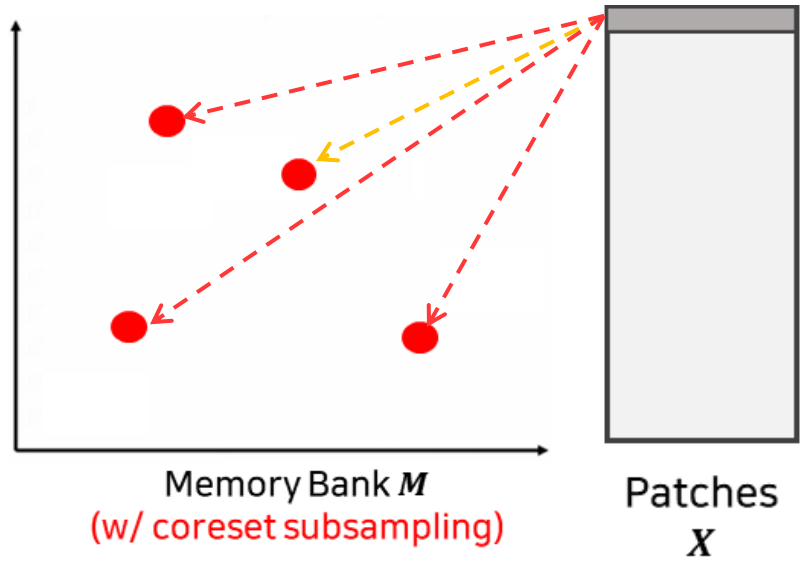
$$x^*, m^* = \arg\max_{x \in X} \arg\min_{m \in M} \|x - m\|_2 \quad (8)$$

$$s^* = \|x^* - m^*\|_2 \quad (9)$$

$$LAS = \delta_1 \cdot s^*_{spatial} + \delta_2 \cdot s^*_{temporal} \quad (10)$$

$$GAS = s^*_{high-level} \quad (11)$$

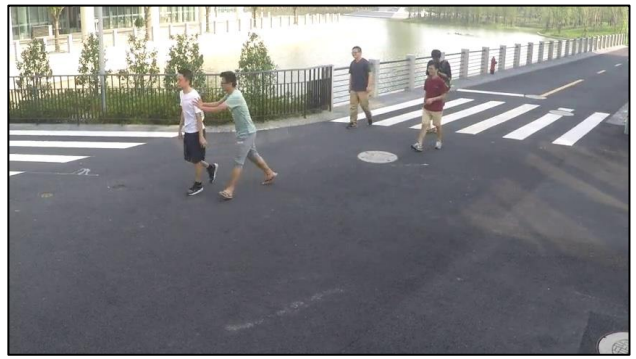
$$Anomaly\ Score = \mathfrak{r}_1 \cdot \frac{LAS - \mu(LAS)}{\sigma(LAS)} + \mathfrak{r}_2 \cdot \frac{GAS - \mu(GAS)}{\sigma(GAS)} \quad (12)$$



04 Experiments Datasets

CUHK Avenue (Avenue)

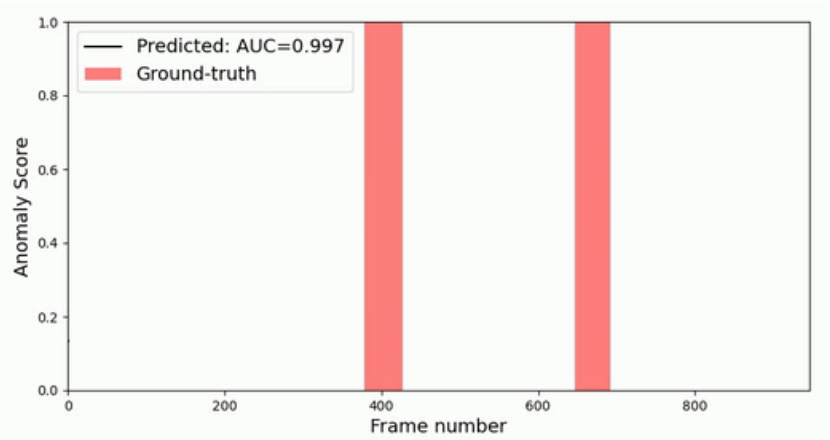
ShanghaiTech (SHTech)



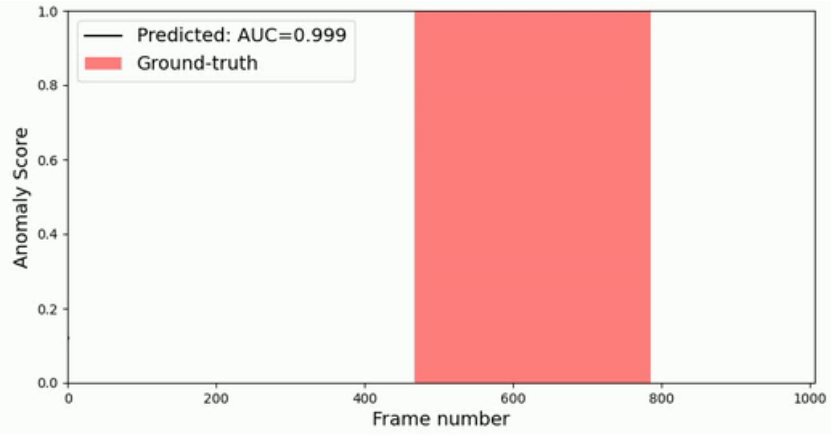
IITB Corridor (Corridor)

Dataset	Training frames (normal)	Testing frames	Abnormal events
Avenue	15,328	15,324	Running, Throwing object, Wrong direction, etc
SHTech	2,74,515	42,883	Throwing object, Jumping, Pushing, Riding a bike, Climbing, etc
Corridor	3,01,999	1,81,567	Protest, Unattached Baggage, Cycling, Sudden Running, Fighting, Playing with Ball, etc

04 Experiments Demo



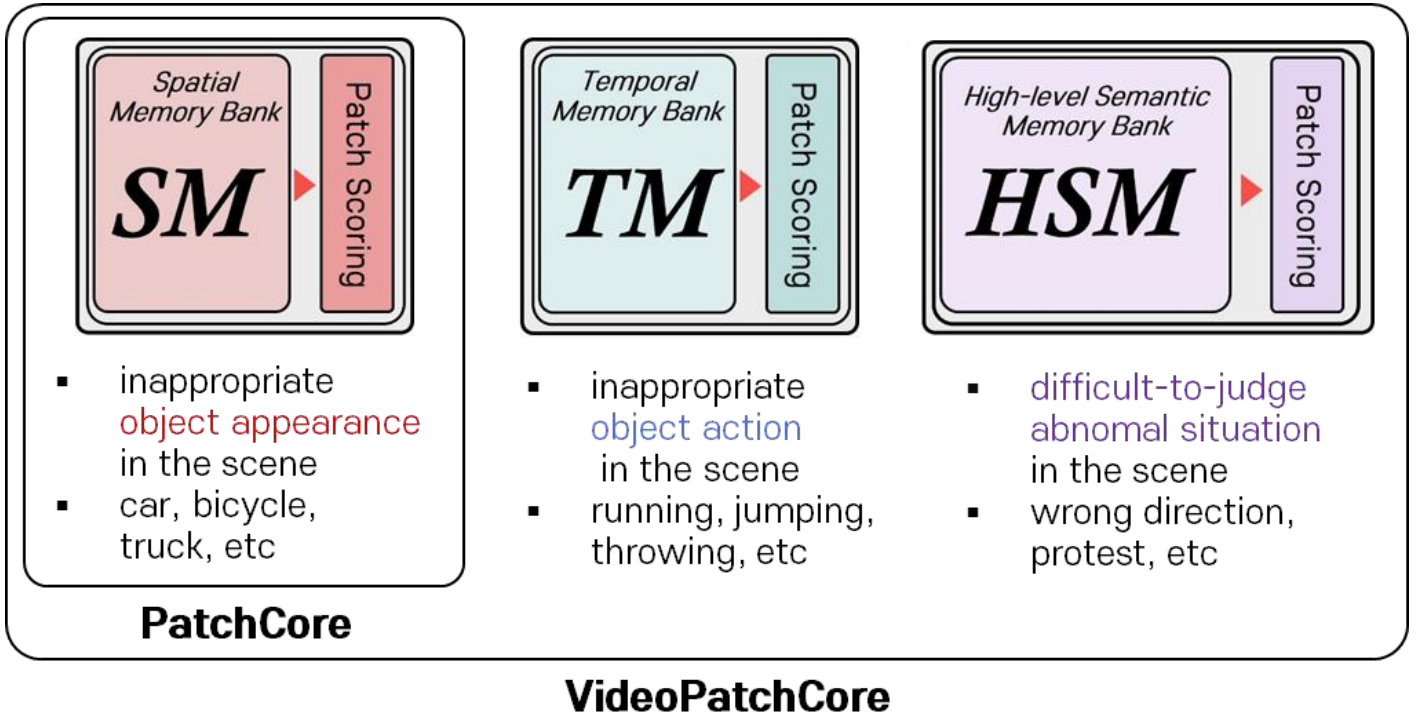
- A video of an abnormal event where a person is **running** on campus.



- A video of an abnormal event where a person is **throwing a bag** on campus.

04 Experiments Qualitative Evaluation

- **PatchCore** detects anomalies through the appearance information of objects.
- **VideoPatchCore** detects anomalies by utilizing the appearance and motion information of objects, along with high-level information from frames.



04 Experiments Qualitative Evaluation

- **Anomaly Score Visualization *[Avenue]***

(a) A person throwing paper

(b) Wrong direction

- **Necessitate consideration of motion information.** Temporal memory plays a crucial role in this scenario.
- **Necessitate consideration of both object and scene contexts.** High-level semantic memory plays a crucial role in this scenario.

04 Experiments Qualitative Evaluation

- **Anomaly Score Visualization** *[SHTech]*

(a) Skateboarding

(b) Fighting

- **Necessitate consideration of motion information.** Temporal memory plays a crucial role in this scenario.
- **Necessitate consideration of interactions between two peoples.** High-level semantic memory plays a crucial role in this scenario.

04 Experiments Qualitative Evaluation

- **Anomaly Score Visualization** ***[Corridor]***

(a) Sudden running

(b) Suspicious object

- **Necessitate consideration of motion information.** Temporal memory plays a crucial role in this scenario.
- **Necessitate consideration of relationship between the person and object.** High-level semantic memory plays a crucial role in this scenario.

04 Experiments Ablation Study (Spatial and Temporal memory)

- Spatial Memory allows for the effective detection of **appearance anomaly events (i.e., cars, bicycles)**.
- Temporal Memory allows for the effective detection of **motion anomaly events (i.e., jumping, running)**.
- By utilizing both Spatial and Temporal Memory, anomaly detection can be performed by considering both appearance and motion status.

Table 2. Comparison of the AUROC scores for the spatial and temporal memory on the Avenue, SHTech and Corridor datasets.

Method	Avenue	SHTech	Corridor
Spatial	84.8%	74.7%	70.5%
Temporal	66.9%	78.8%	73.5%
Spatial+Temporal	**90.3%**	**84.8%**	**76.3%**

04 Experiments Ablation Study (Split pooling)

- **Head** compresses lf by selecting c' channels from the front.
- **Random** compresses lf by selecting random c' channels.
- **Split pool** divides lf into c' groups and compresses by averaging the channels of each group.
- Proposed method achieves high performance by effectively compressing the original information.

Table 3. Comparison of the AUROC scores for compression methods on the Avenue, SHTech and Corridor datasets.

Compression	Avenue	SHTech	Corridor
Head	82.7%	76.9%	74.9%
Random	85.9%	83.8%	74.8%
Split Pool	**90.3%**	**84.8%**	**76.3%**

04 Experiments Ablation Study (Mismatched pooling)

- **Avg Pool** uses GAP in both the memorization and inference stages, **Max Pool** applies GMP in both stages.
- **Mismatched Pool** performs GAP during the memorization and GMP during the inference stage.
- Differentiating between normal and abnormal temporal patches
  by varying the pooling methods at each stage is effective for VAD.

Table 4. Comparison of AUROC scores for feature pooling methods in the temporal partition across the Avenue, SHTech and Corridor datasets.

Pooling	Avenue	SHTech	Corridor
Avg Pool	84.9%	76.1%	74.7%
Max Pool	88.2%	83.8%	75.7%
Mismatched Pool	**90.3%**	**84.8%**	**76.3%**

04 Experiments Ablation Study (Local and Global stream)

- Using the **local stream** complicates detecting scene anomalies or abnormal interactions between multiple objects.
- Using the global stream together enhances VAD by leveraging broader spatiotemporal information.
- In the latter datasets, there are many situations where objects are adjacent to each other, allowing the local stream alone to partially fulfill the role of the global stream.

Table 5. Comparison of the AUROC scores for the local and global stream on the Avenue, SHTech and Corridor datasets.

Stream	Avenue	SHTech	Corridor
Local	90.3%	84.8%	76.3%
Global	84.4%	68.4%	67.2%
Local+Global	**92.8%**	**85.1%**	**76.4%**

04 Experiments Further Analysis (Coreset subsampling ratio)

- Higher subsampling rates store more normal features in memory, improving performance but slowing down speed.
- However, using Coreset subsampling results in a minimal performance difference of 1% and 99% (less than 0.5%).
- No need to consider the memory size for large video datasets.

Table 6. Comparison of AUROC scores for the subsampling ratio on the SHTech and Corridor datasets.

Subsampling ratio	SHTech		Corridor	
	AUC	FPS	AUC	FPS
1%	84.6% (-0.5%)	**170.9**	76.0% (-0.4%)	**143.4**
10%	85.0% (-0.1%)	154.8	**76.4% (-0.0%)**	113.5
25%	**85.1% (-0.0%)**	96.1	76.3% (-0.1%)	60.0
50%	**85.1% (-0.0%)**	54.8	76.3% (-0.1%)	40.0
75%	**85.1% (-0.0%)**	39.6	76.3% (-0.1%)	25.2
99%	**85.1%**	31.0	**76.4%**	19.5

04 Experiments Further Analysis (Memorizing Technique)

- Former method constructs memory from each video and concatenates them.
- Latter method constructs memory from all videos collectively.
- Expected that former method would be less constrained by hardware limitations
  but might suffer from a lower performance due to the uneven distribution of features.
- However, the performance difference between the two methods is almost nonexistent.

Table 7. Comparison of AUROC scores for the memorizing technique at memory usage levels of both10% and 99% on the SHTech and Corridor datasets.

Memorizing technique	Ratio	SHTech	Corridor
Subsampling → concat	10%	85.0%	**76.4%**
	99%	**85.1%**	**76.4%**
Concat → Subsampling	10%	84.9% (-0.1%)	76.3% (-0.1%)
	99%	85.0% (-0.1%)	**76.4% (-0.0%)**

04 Experiments Quantitative Evaluation

- **Comparison with SOTA methods**

Method	Venue	Memory	Avenue	SHTech	Corridor
FFP [18]	CVPR 18		84.9%	72.8%	64.7%
MPED-RNN [24]	CVPR 19		-	73.4%	64.3%
MemAE [8]	ICCV 19	✓	83.3%	71.2%	-
AMC [25]	ICCV 19		86.9%	-	-
MTP [29]	WACV 20		82.9%	76.0%	67.1%
CDDA [5]	ECCV 20		86.0%	73.3%	-
MNAD [26]	CVPR 20	✓	88.5%	70.5%	-
ROADMAP [34]	TNNLS 21		88.3%	76.6%	-
AMMC-Net [3]	AAAI 21	✓	86.6%	73.7%	-
MPN [23]	CVPR 21	✓	89.5%	73.8%	-
HF^2-VAD [20]	ICCV 21	✓	91.1%	76.2%	-
LLSH [22]	TCSVT 22		87.4%	77.6%	73.5%
VABD [17]	TIP 22		86.6%	78.2%	72.2%
DLAN-AC [37]	ECCV 22	✓	89.9%	74.7%	-
Jigsaw [33]	ECCV 22		92.2%	84.3%	-
Sun et al. [31]	AAAI 23	✓	91.5%	78.6%	-
Cao et al. [4]	CVPR 23		86.8%	79.2%	73.6%
USTN-DSC [36]	CVPR 23		89.9%	73.8%	-
DMAD [19]	CVPR 23	✓	**92.8%**	78.8%	-
FPDM [35]	ICCV 23		90.1%	78.6%	-
STG-NF* [11]	ICCV 23		61.8%	**85.9%**	61.4%
HSC [32]	ICCV 23	✓	92.4%	83.0%	-
Ristea et al. [28]	CVPR 24		91.3%	79.1%	-
Zhang et al. [39]	CVPR 24		92.4%	85.1%	-
VPC (Ours)	-	✓	**92.8%**	85.1%	**76.4%**

- It demonstrates the competitive performance compared to other state-of-the-art (SOTA) methods
- Compared to the other methods using memory, a superior performance is achieved.
- When compared to HSC, which utilize appearance and motion memory, our approach outperforms by 0.4% and 2.1% on the Avenue and SHTech datasets, respectively, leveraging three memory components effectively.

05 Conclusion Contributions

- I propose VPC, an extension of PatchCore developed for image anomaly detection,
  to perform effective video anomaly detection.

- VPC employs two streams (local and global) and three memory banks (spatial, temporal, and high-level semantic)
  to capture the spatiotemporal characteristics of videos and detect various forms of anomalies.

- VPC achieves good performance comparable to state-of-the-art methods.

05 Conclusion Future works

- I will utilize text embeddings obtained from language models to enable High-level Semantic Memory
  to more clearly understand the relationships between objects.

- Current memory is simply designed without training, making it challenging to accommodate various situations.
  Since memory needs to be constructed differently based on situations, I will develop a robust memory for diverse situations.

- I will develop an efficient deep learning network and memory to perform
  real-time video anomaly detection in an on-device environment.

06 Publications

- **Sunghyun Ahn**, Youngwan Jo, Kijung Lee, and Sanghyun Park., "VideoPatchCore: An Effective Method to Memorize Normality for Video Anomaly Detection", Asian Conference on Computer Vision (ACCV), 2024.

- Seungkyun Hong*, **Sunghyun Ahn***, Youngwan Jo, and Sanghyun Park. (*equally contributed), "Making Anomalies More Anomalous: Video Anomaly Detection Using a Novel Generator and Destroyer", IEEE Access, 2024.

- Seungkyun Hong*, **Sunghyun Ahn***, Youngwan Jo, and Sanghyun Park. (*equally contributed), "Dual Stream Fusion U-Net Transformers for 3D Medical Image Segmentation", IEEE International Conference on Big Data and Smart Computing (BigComp), 2024.

Thank you

07 Appendix Quantitative Evaluation

- **Detailed Analysis of Subsampling Ratio**

Avenue	1%	10%	25%	50%	75%	99%
Spatial	0.848	0.831	0.828	0.825	0.828	0.828
Temporal	0.669	0.669	0.669	0.669	0.669	0.669
High-level	0.844	0.845	0.848	0.844	0.844	0.844
Total	0.928	0.918	0.914	0.912	0.912	0.912

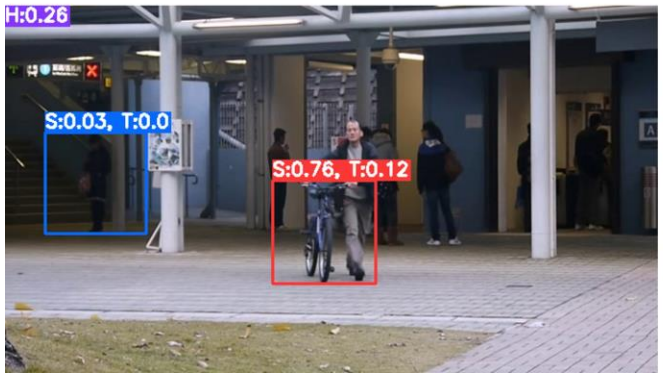
SHTech	1%	10%	25%	50%	75%	99%
Spatial	0.748	0.744	0.747	0.747	0.746	0.746
Temporal	0.788	0.788	0.788	0.788	0.788	0.788
High-level	0.671	0.675	0.684	0.673	0.673	0.674
Total	0.846	0.850	0.851	0.851	0.851	0.851

Corridor	1%	10%	25%	50%	75%	99%
Spatial	0.690	0.705	0.705	0.705	0.706	0.705
Temporal	0.735	0.735	0.735	0.735	0.735	0.735
High-level	0.664	0.672	0.673	0.674	0.675	0.660
Total	0.760	0.764	0.763	0.763	0.763	0.764

- The performance difference between using 10% and 99% of memory is very small.
- In practical use, sufficiently good performance can be maintained even with memory usage set at 10% or lower.

07 Appendix Qualitative Evaluation

- **Object-wise Anomaly Scores *[Avenue]***



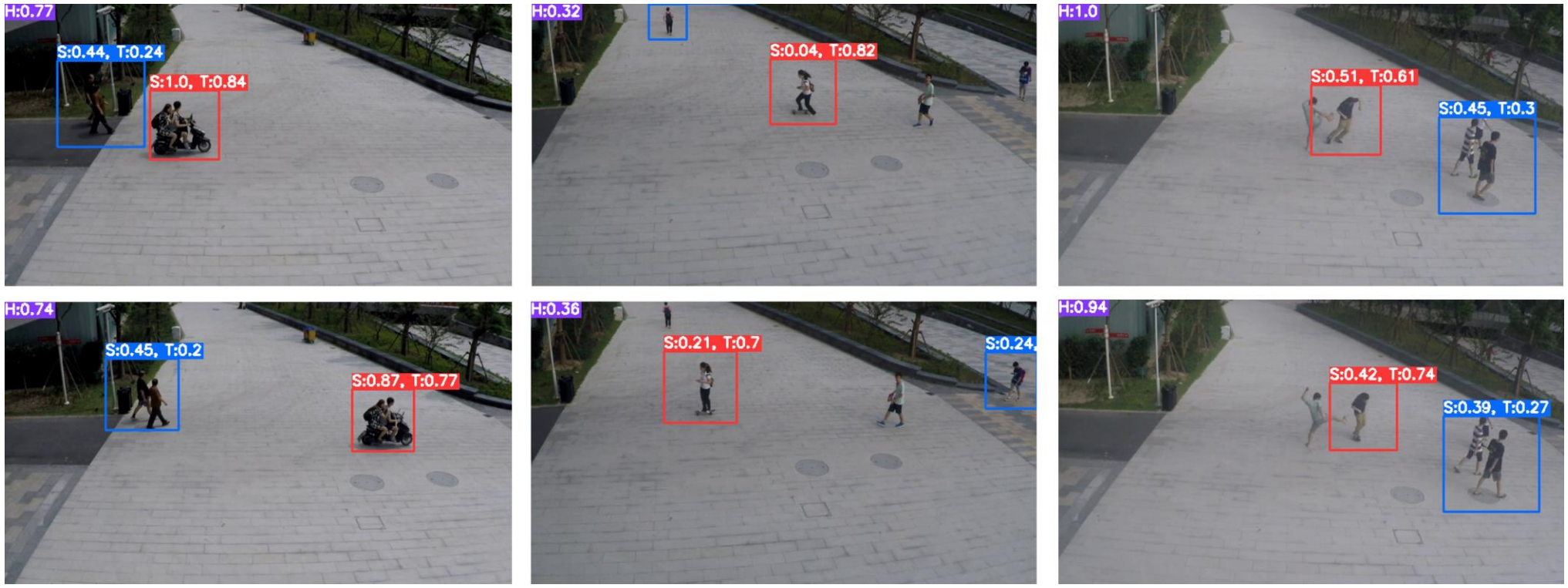
Bicycle

Running

Wrong direction

07 Appendix Qualitative Evaluation

- **Object-wise Anomaly Scores *[SHTech]***



Motorcycle Skateboarding Fighting

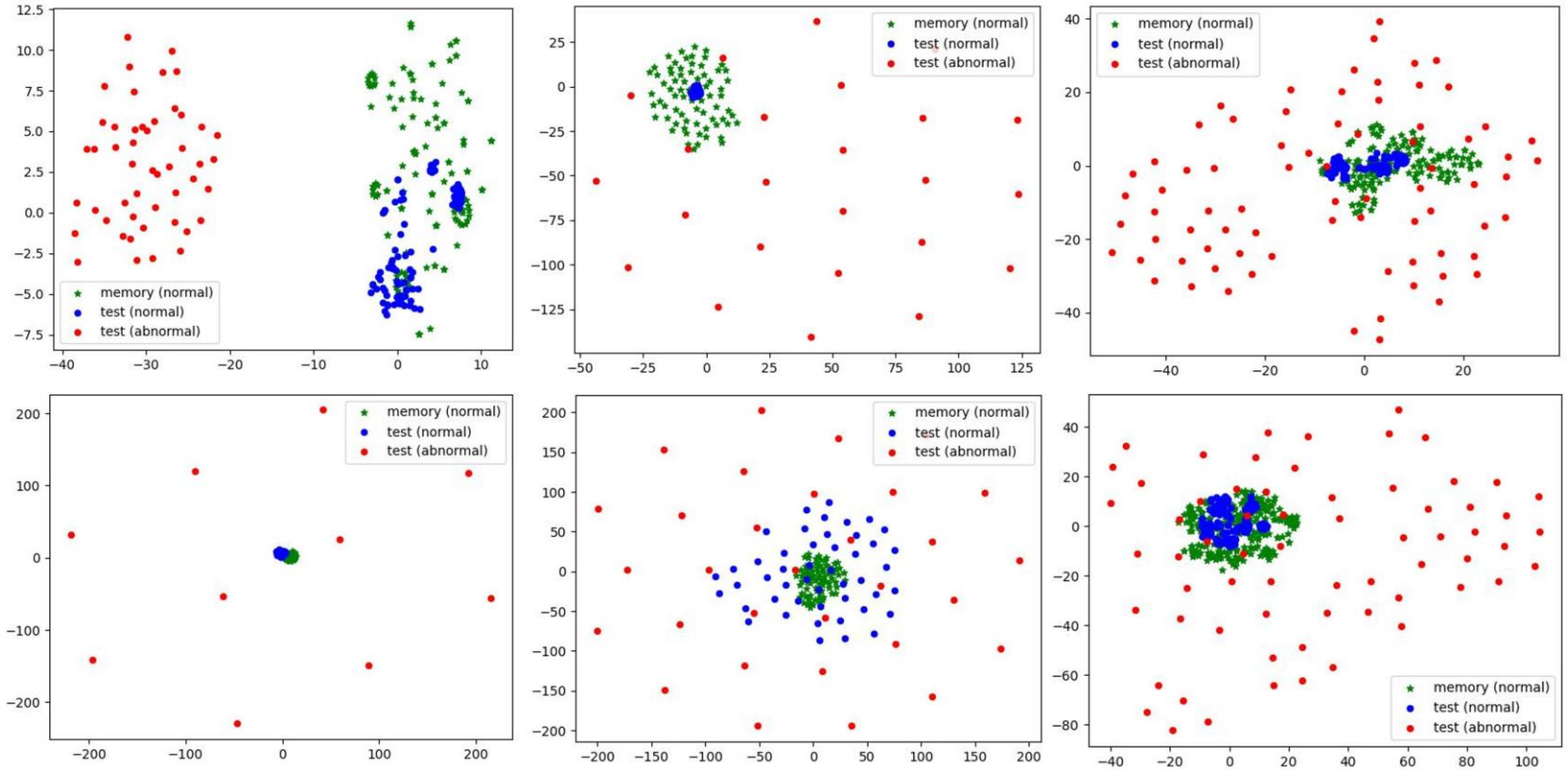
07 Appendix Qualitative Evaluation

- **Object-wise Anomaly Scores *[Corridor]***

Suspicious object | Playing with ball | Protesting

07 Appendix Qualitative Evaluation

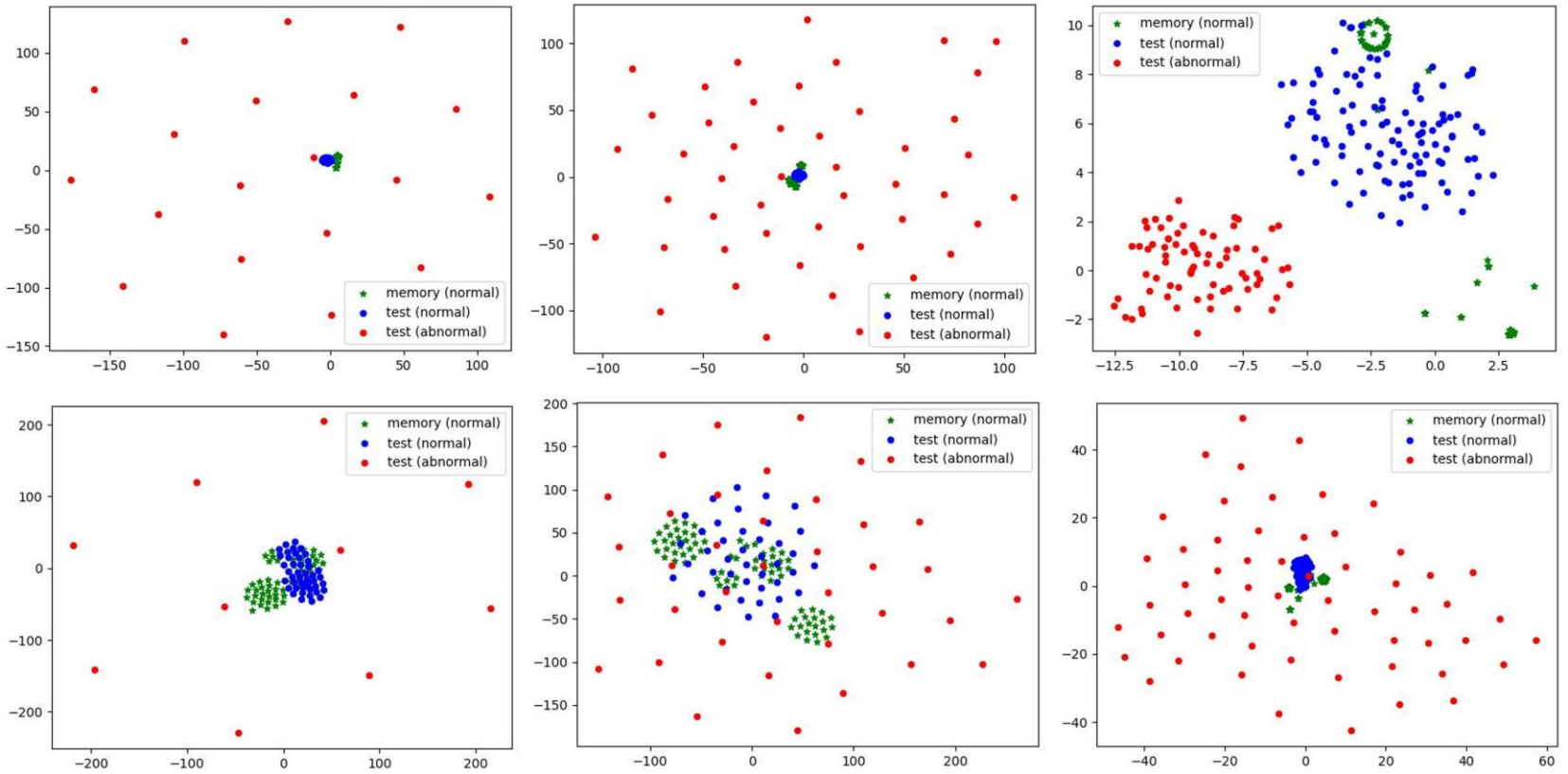
- **t-SNE Visualization of Memory and Test Patches *[Spatial]***



Avenue – 01, 03 SHTech – 18, 28 Corridor – 09, 35

07 Appendix Qualitative Evaluation

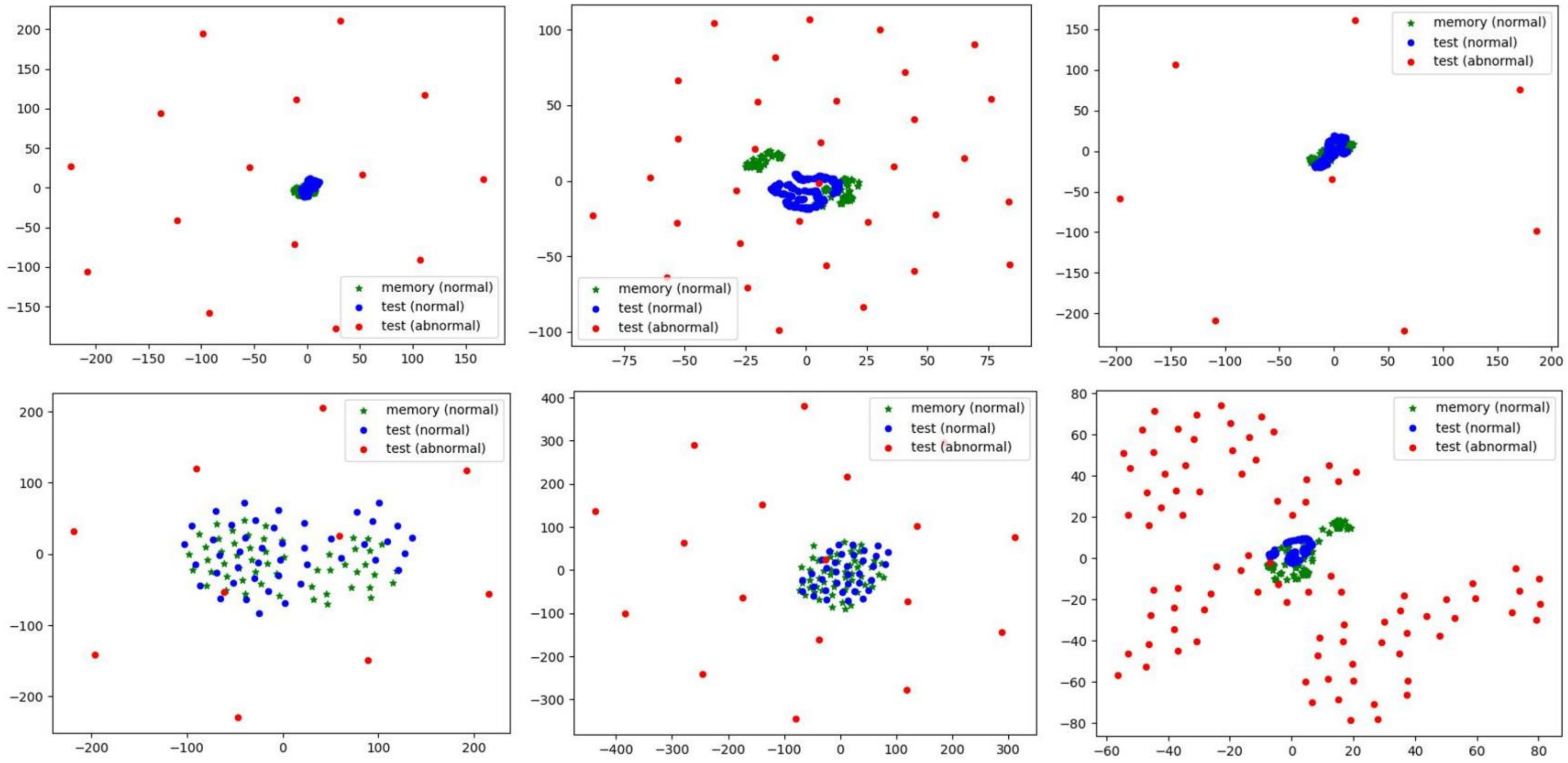
- **t-SNE Visualization of Memory and Test Patches *[Temporal]***



Avenue – 10, 13 SHTech – 24, 86 Corridor – 45, 140

07 Appendix Qualitative Evaluation

- **t-SNE Visualization of Memory and Test Patches** *[High-level]*



Avenue – 02, 13 | SHTech – 49, 68 | Corridor – 18, 85