Video Anomaly Detection

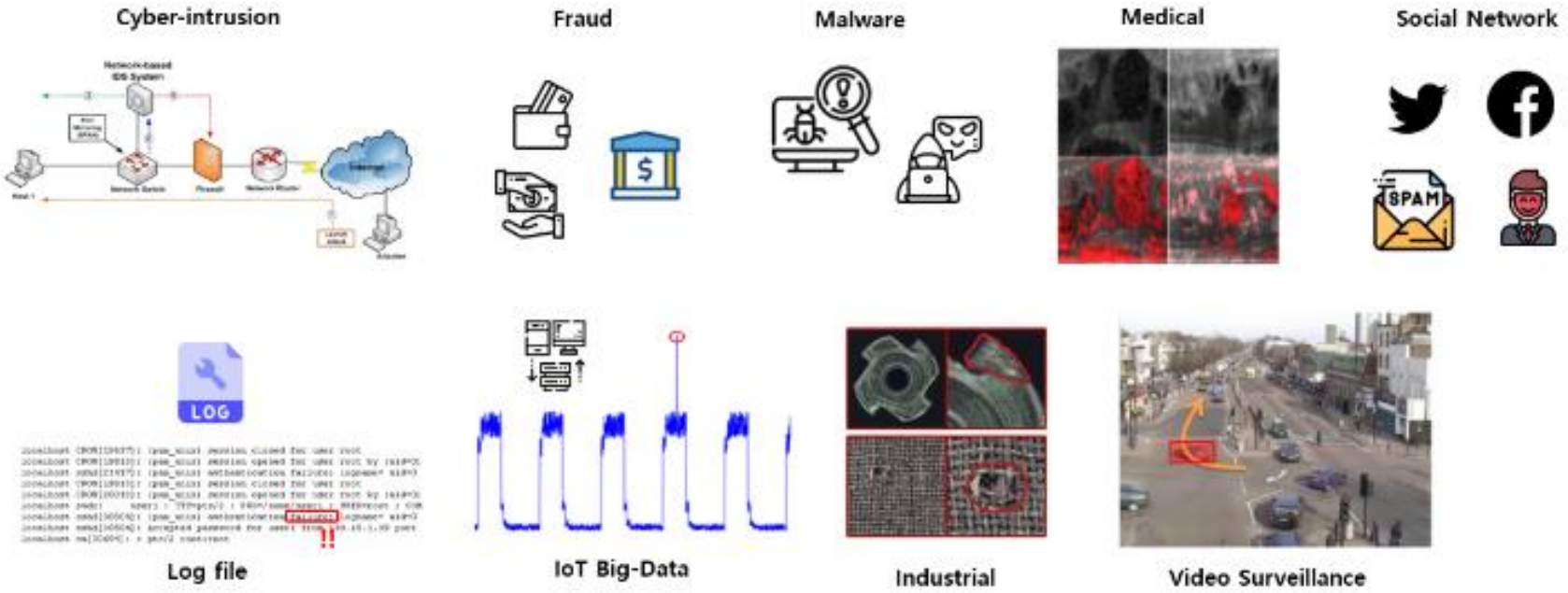
Application In Database Systems
(CSI8782.01-01)

Data Engineering Lab
Multi Modal Deep Learning Team

연세대학교

Anomaly Detection

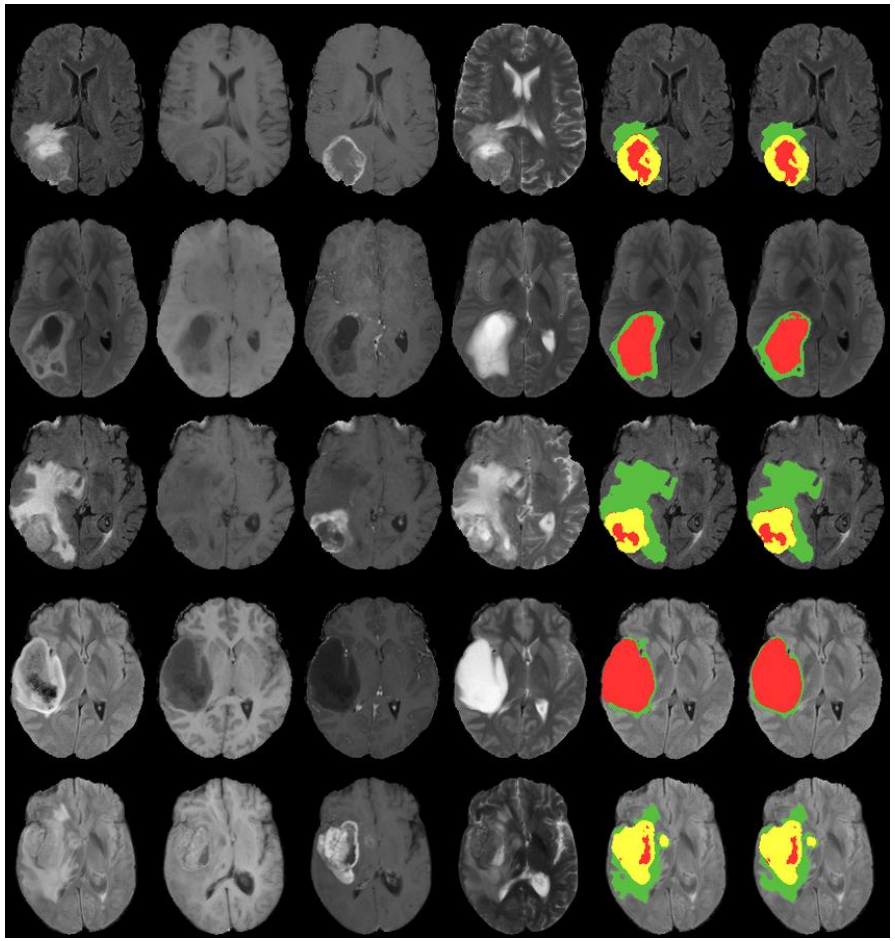
- **What is Anomaly Detection?**
  - ✓ Anomaly Detection is a task that distinguishes between normal and abnormal samples.
  - ✓ Animal Detection is used in various fields such as time series data, images, and etc.



https://hoya012.github.io/blog/anomaly-detection-overview-1/

Anomaly Detection

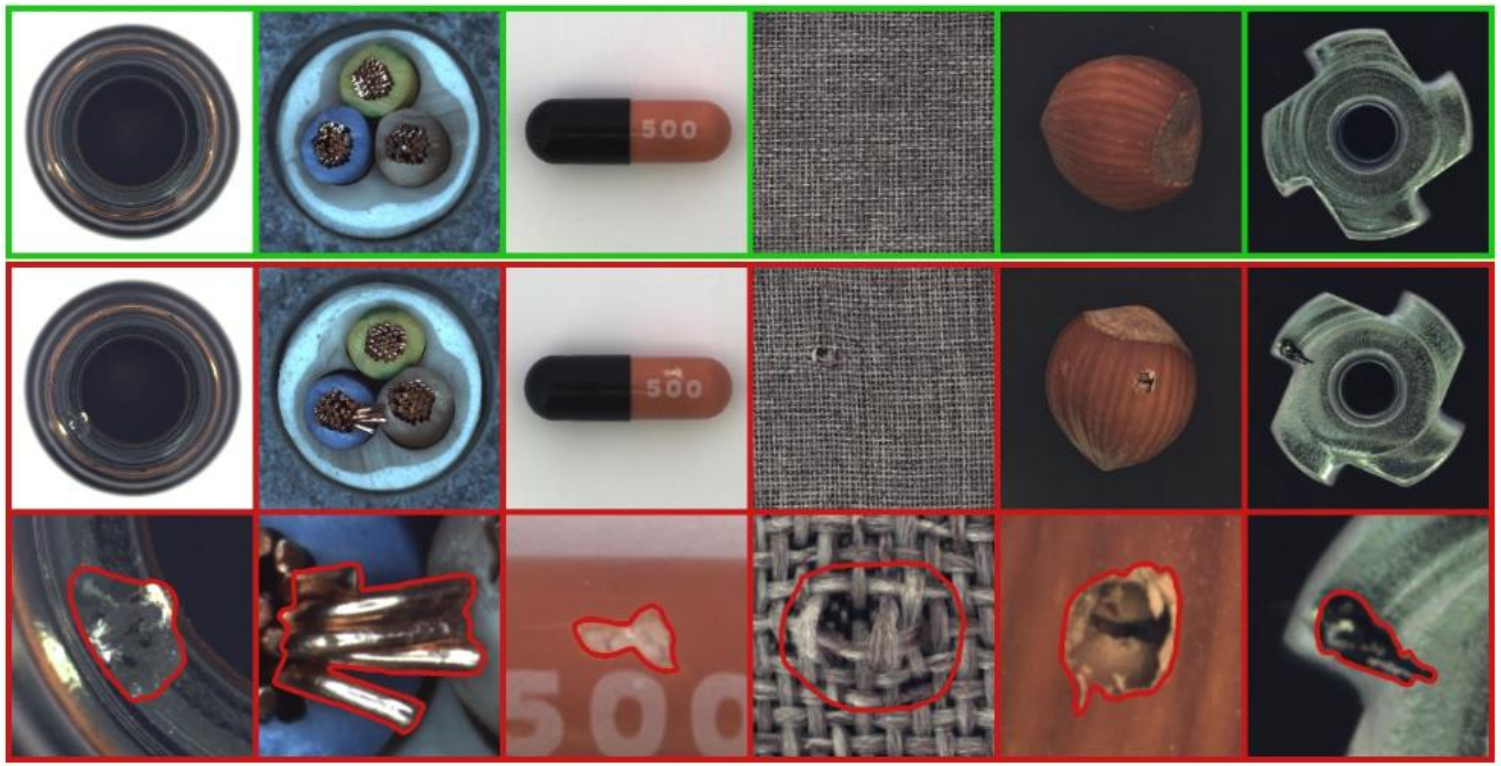
- **Example of Anomaly**
  - ✓ Brain Tumor



BraTS 2019 Dataset

Anomaly Detection

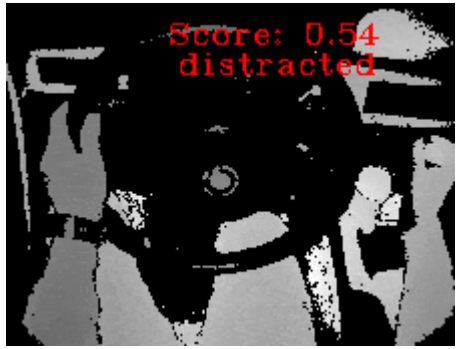
- **Example of Anomaly**
  - ✓ Defective products

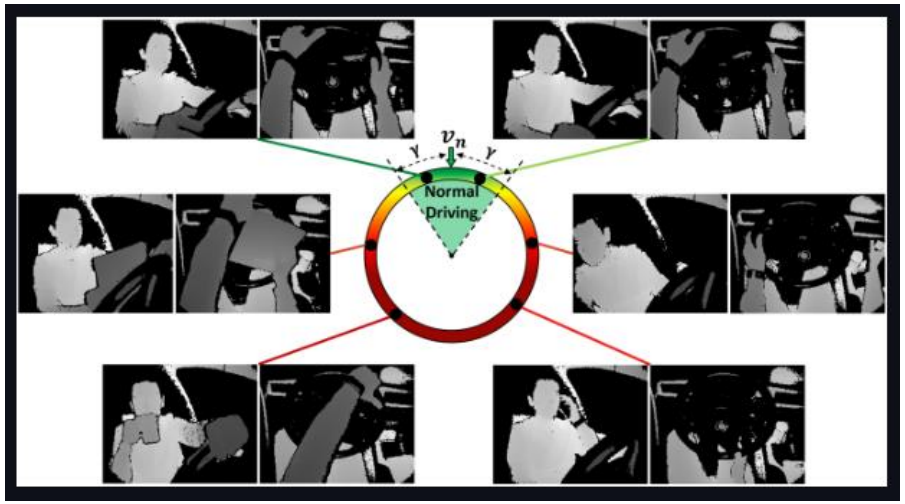


MVTec Anomaly Detection Dataset

Anomaly Detection

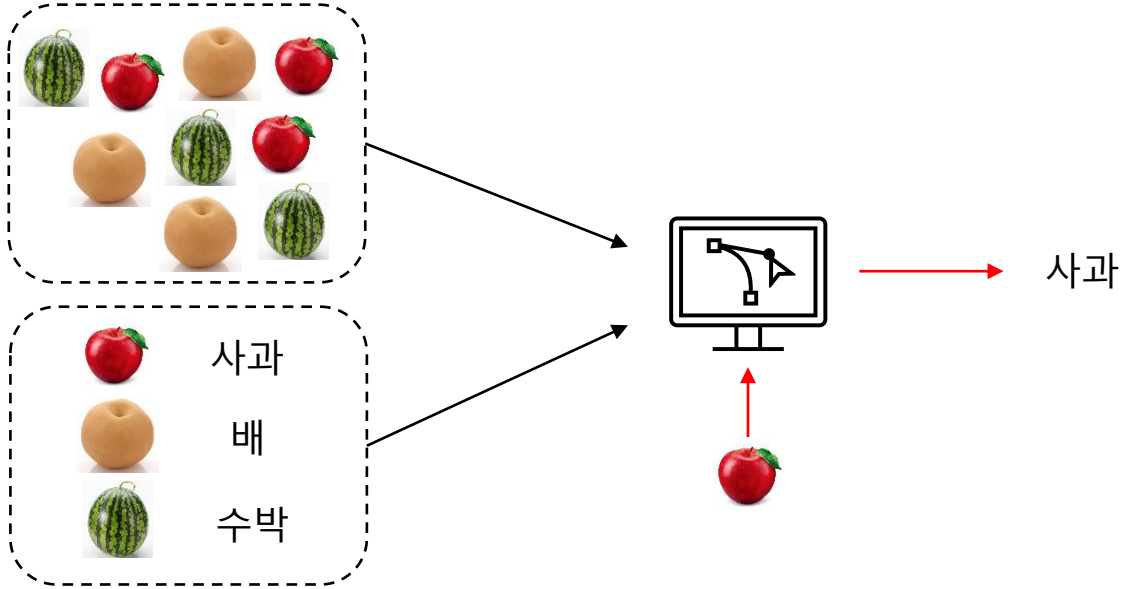
- **Example of Anomaly**
  - ✓ Driving

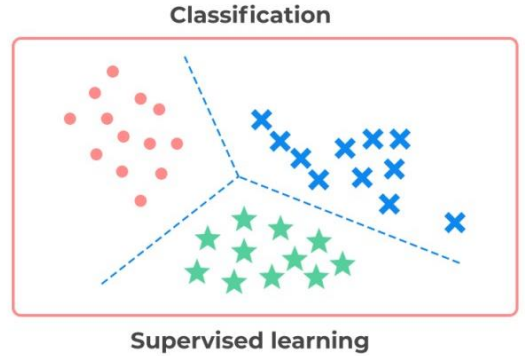


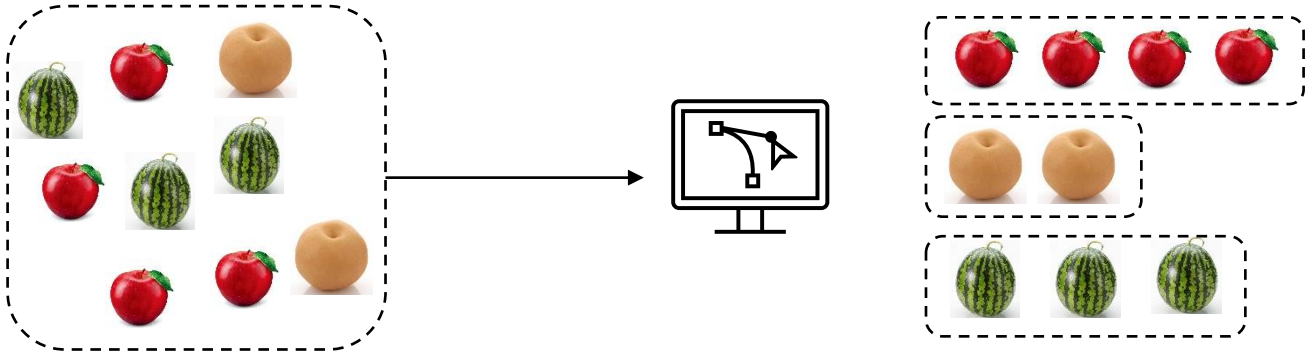


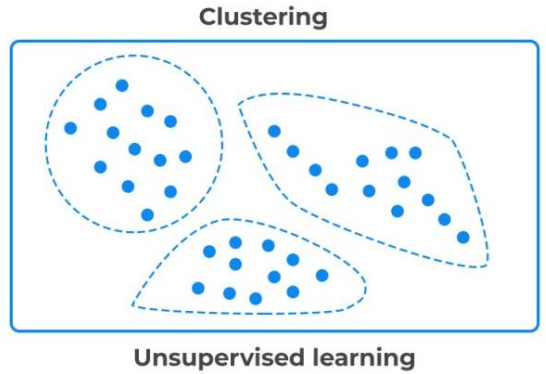
Anomaly Detection

- Supervised vs Unsupervised





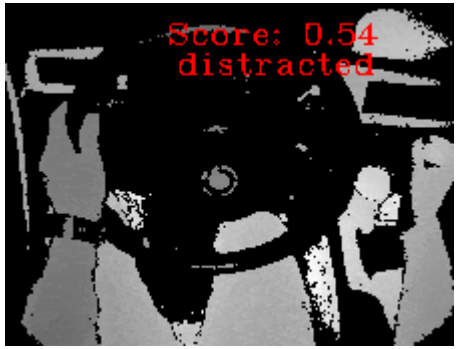




Anomaly Detection

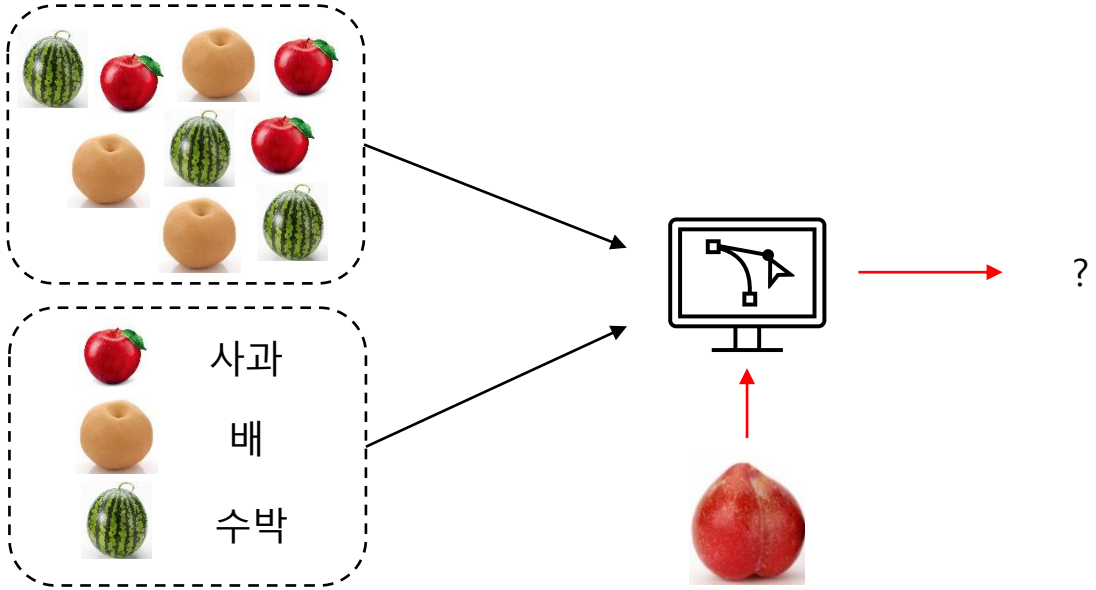
- **Not enough abnormal samples**
  - ✓ Because it is difficult to define anomaly, it is impossible to divide the classes.
  - ✓ Abnormal samples occur very rarely, making it difficult to gather data, which leads to class imbalance problem.
  - ✓ Therefore, unsupervised learning is mainly used rather than supervised learning.

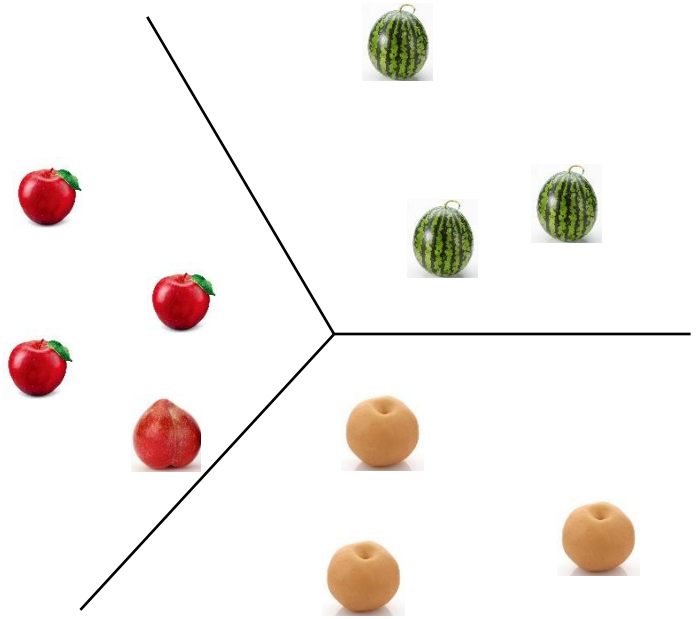
- ✓ Walking person + sidewalk > ?
- ✓ Walking person + driveway > ?
- ✓ Walking person + crosswalk > ?
- ✓ Walking person + crosswalk + Red Traffic Light > ?



Anomaly Detection

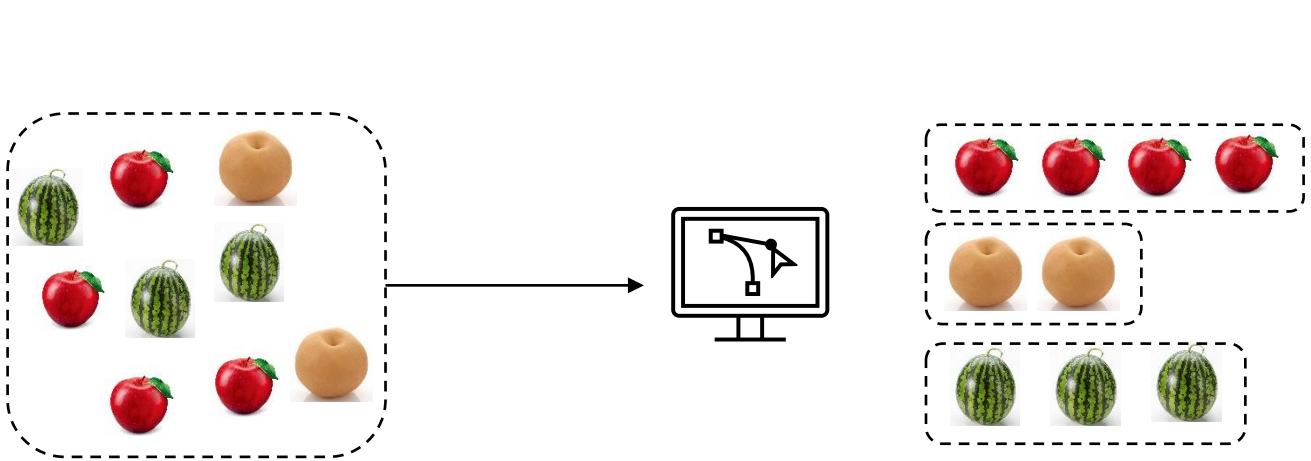
- Supervised vs Unsupervised

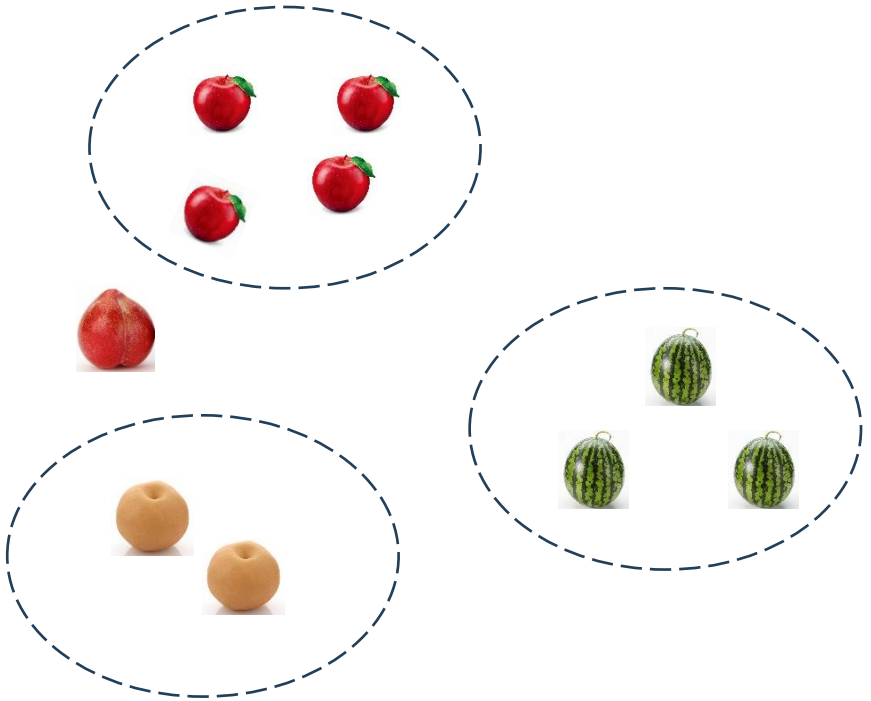




Anomaly Detection

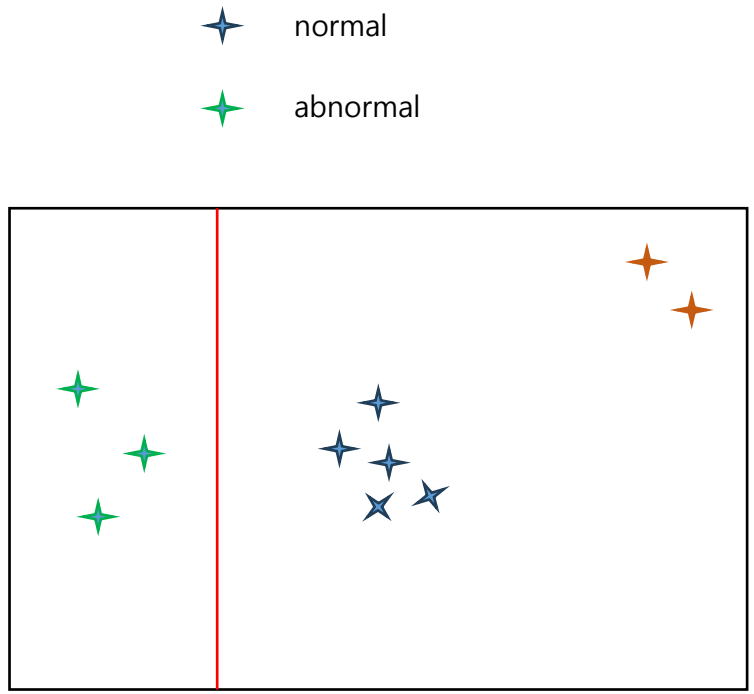
- Supervised vs Unsupervised

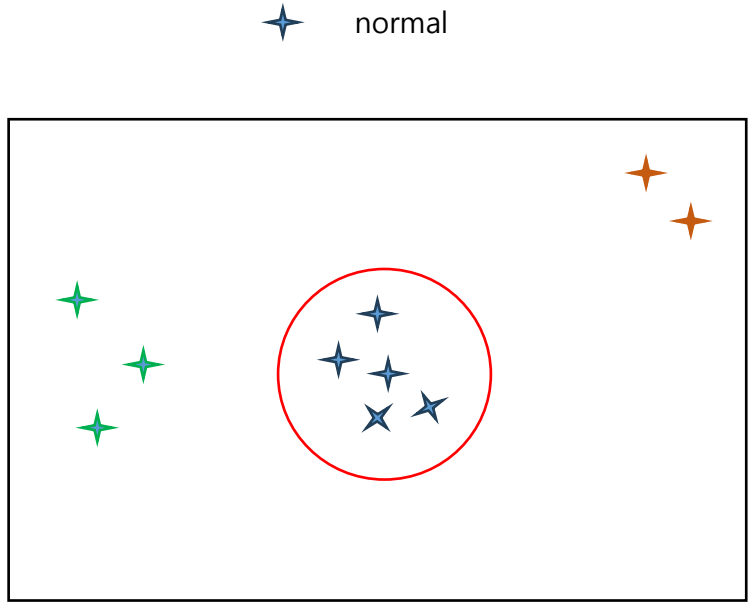




Anomaly Detection

- Supervised vs Unsupervised





Deep One-Class Classification (ICML, 2018)

- **Method**
  - ✓ One-class Method

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

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

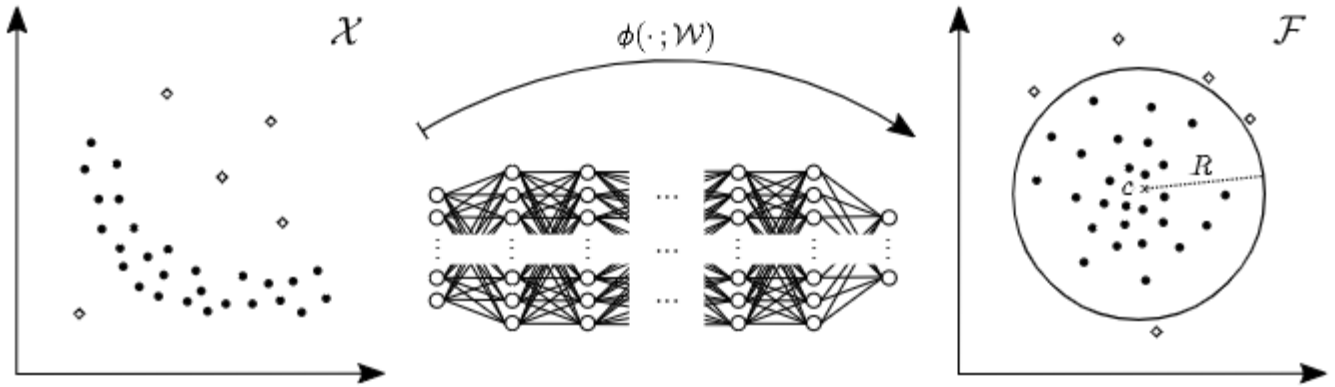
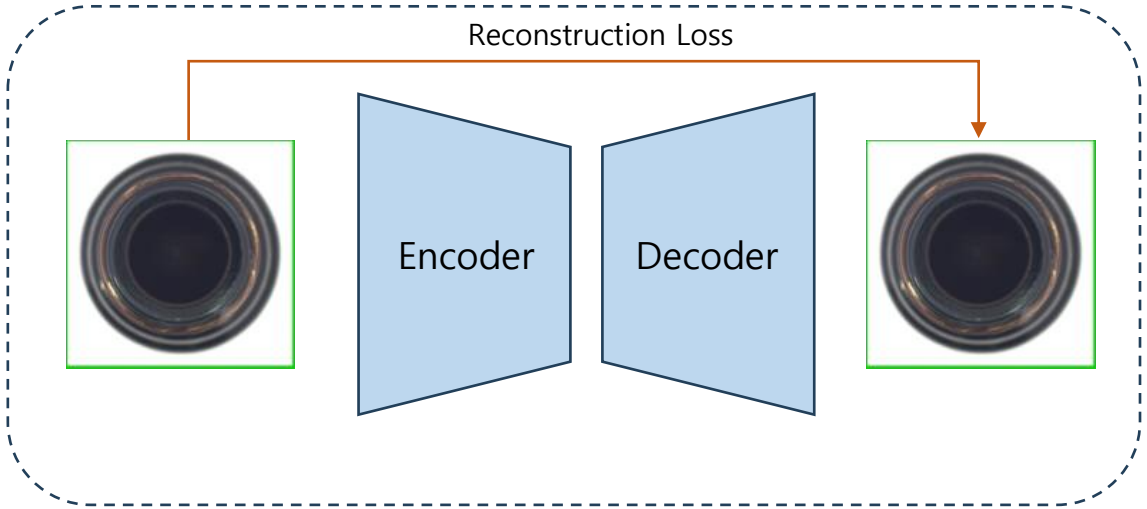


Figure 1. Deep SVDD learns a neural network transformation $\phi(\cdot;\mathcal{W})$ with weights $\mathcal{W}$ from input space $\mathcal{X} \subseteq \mathbb{R}^d$ to output space $\mathcal{F} \subseteq \mathbb{R}^p$ that attempts to map most of the data network representations into a hypersphere characterized by center $\boldsymbol{c}$ and radius R of minimum volume. Mappings of normal examples fall within, whereas mappings of anomalies fall outside the hypersphere.

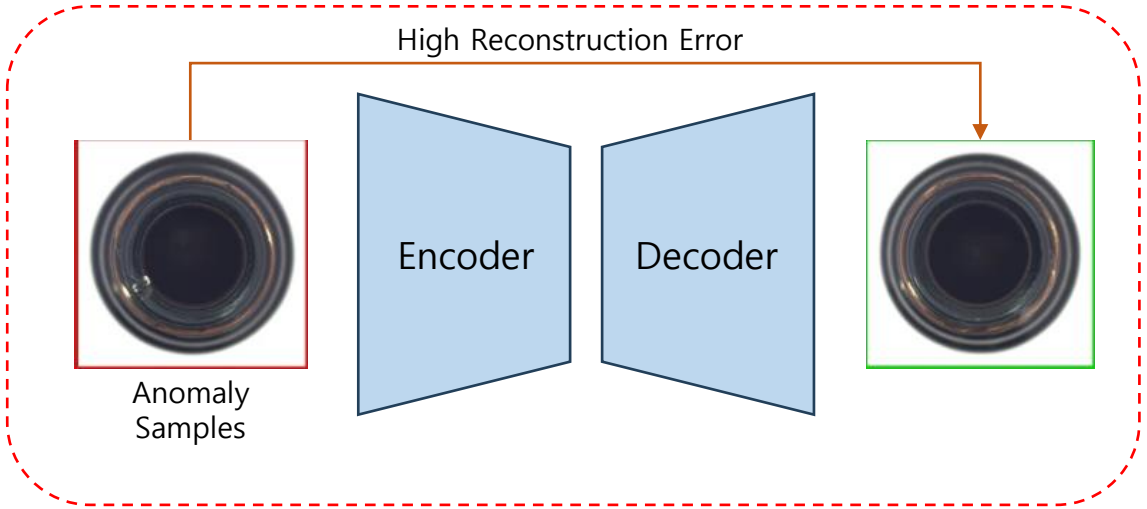
Figure 2. Most normal (left) and most anomalous (right) in-class examples determined by One-Class Deep SVDD for selected MNIST (top) and CIFAR-10 (bottom) one-class experiments.

Anomaly Detection

- **Reconstruction-based Method**
  - ✓ Use an unsupervised method to reconstruct inputs
  - ✓ Only normal samples are used to train and abnormal samples are difficult to reconstruct at test, so determine abnormalities through a large reconstruction error



Training (Only normal Samples)



Test

Unsupervised anomaly detection with generative adversarial networks to guide marker discovery (IPMI, 2017)

- **Method**
  - ✓ One-class (Reconstruction based)
  - ✓ GAN

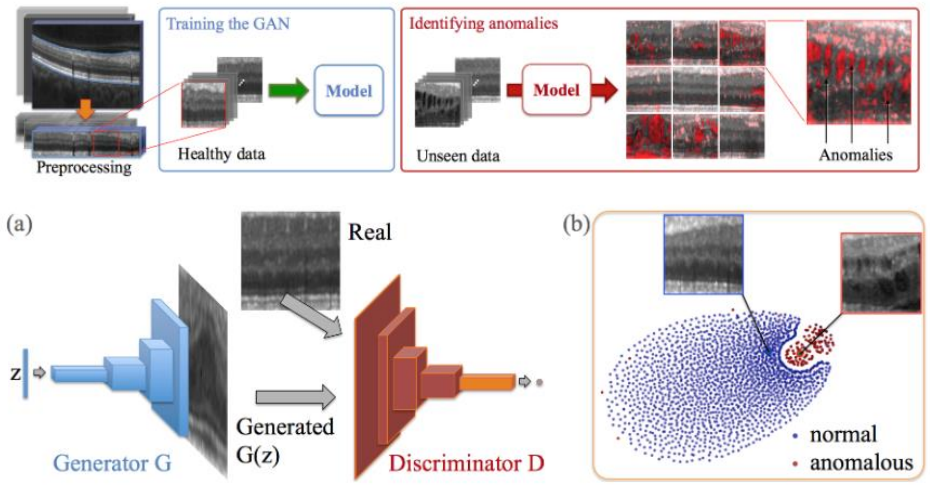
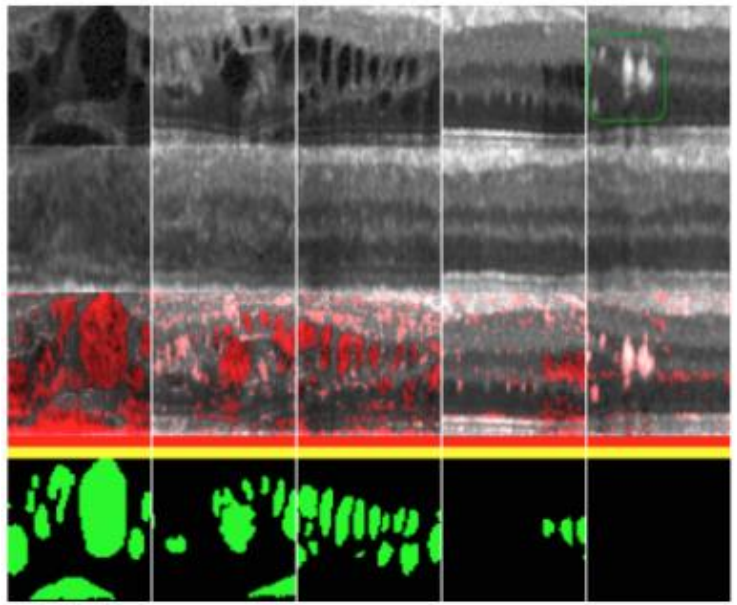


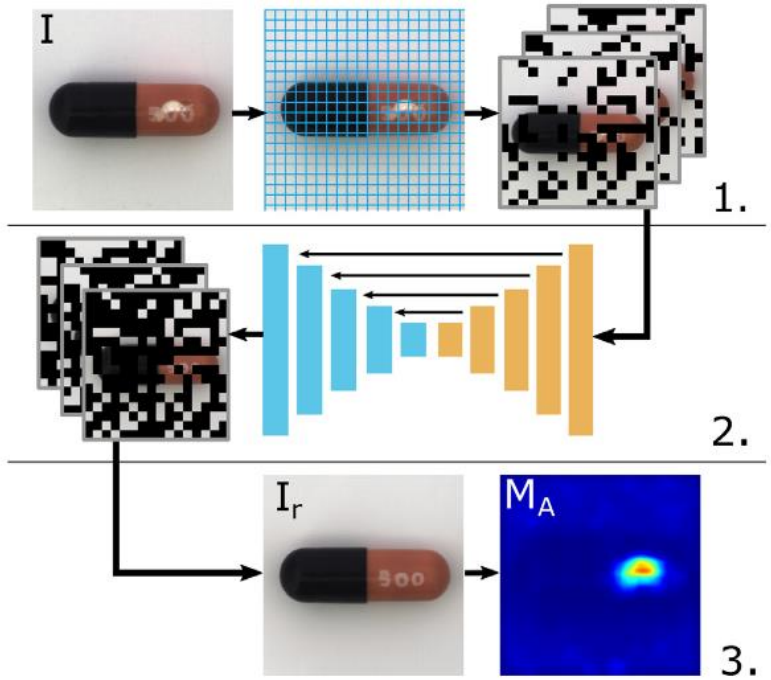
Fig. 2. (a) Deep convolutional generative adversarial network. (b) t-SNE embedding of normal (blue) and anomalous (red) images on the feature representation of the last convolution layer (orange in (a)) of the discriminator.



Schlegl, Thomas, et al. "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery." International conference on information processing in medical imaging. Cham: Springer International Publishing, 2017.

Reconstruction by inpainting for visual anomaly detection (Pattern Recognition, 2021)

- **Method**
  - ✓ One-class (Reconstruction based)
  - ✓ Unet + Inpainting



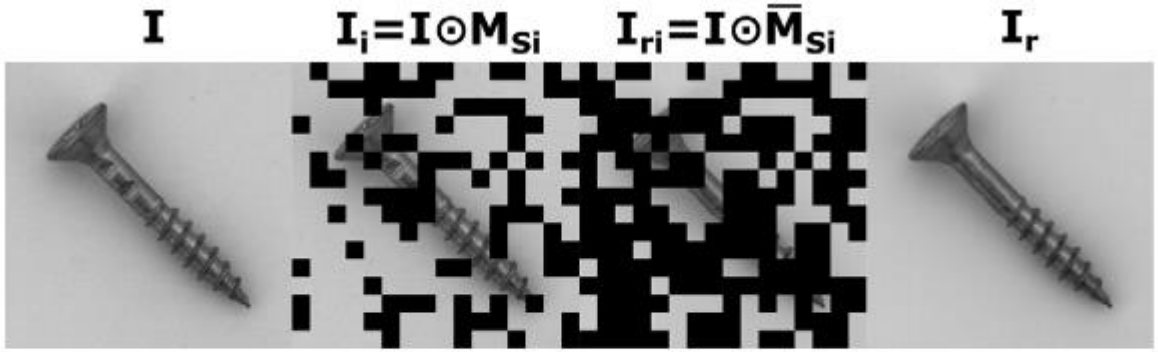
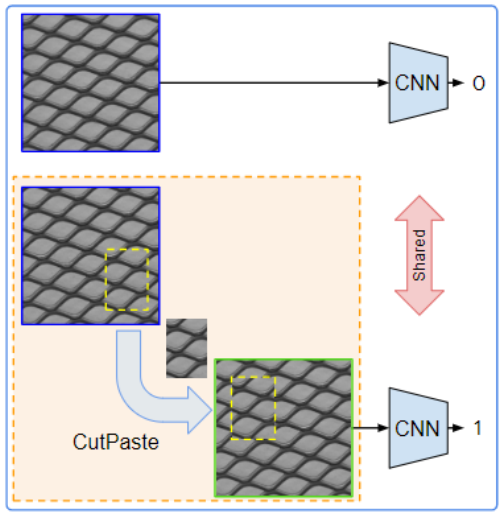


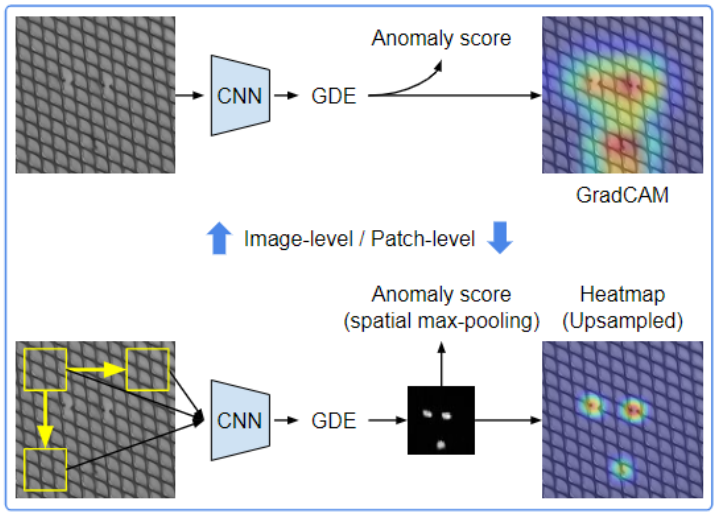
Fig. 3. Regions belonging to S_i are set to zero in image **I** by multiplying **I** by the region mask $\mathbf{M}_{S_i}$. The resulting image $\mathbf{I}_i$ is fed into the network to create a partial reconstruction $\mathbf{I}_{ri}$. The regions in $\mathbf{I}_{ri}$ not belonging to S_i are set to zero by multiplying with the inverse of $\mathbf{M}_{S_i}$. The final image reconstruction $\mathbf{I}_r$ is assembled from partial reconstructions $\mathbf{I}_{ri}$.

Cutpaste: Self-supervised learning for anomaly detection and localization (CVPR, 2021)

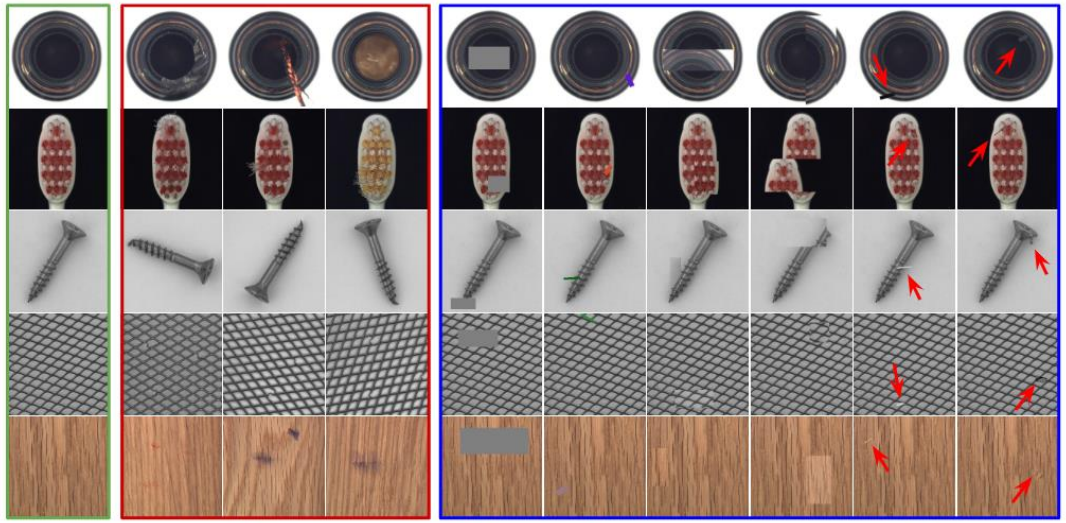
- **Method**
  - ✓ One-class (Self-supervised learning)
  - ✓ Training with self-define anomaly samples by augmentation (e.g. cut, scar, paste)



(a) Learning Self-Supervised Representation



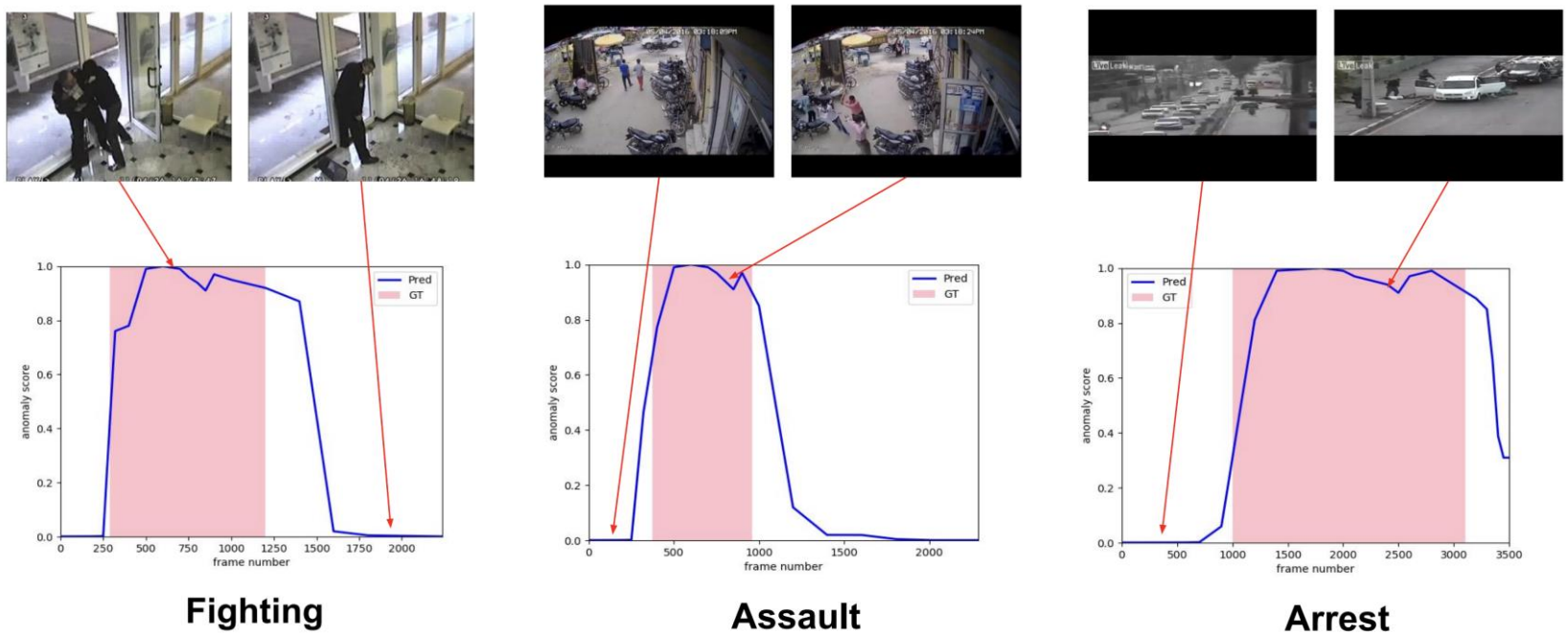
(b) Anomaly Detection and Localization



(a) Normal (b) Anomaly (c) Cutout (d) Scar (e) CutPaste (f) CutPaste (Scar)

Video Anomaly Detection

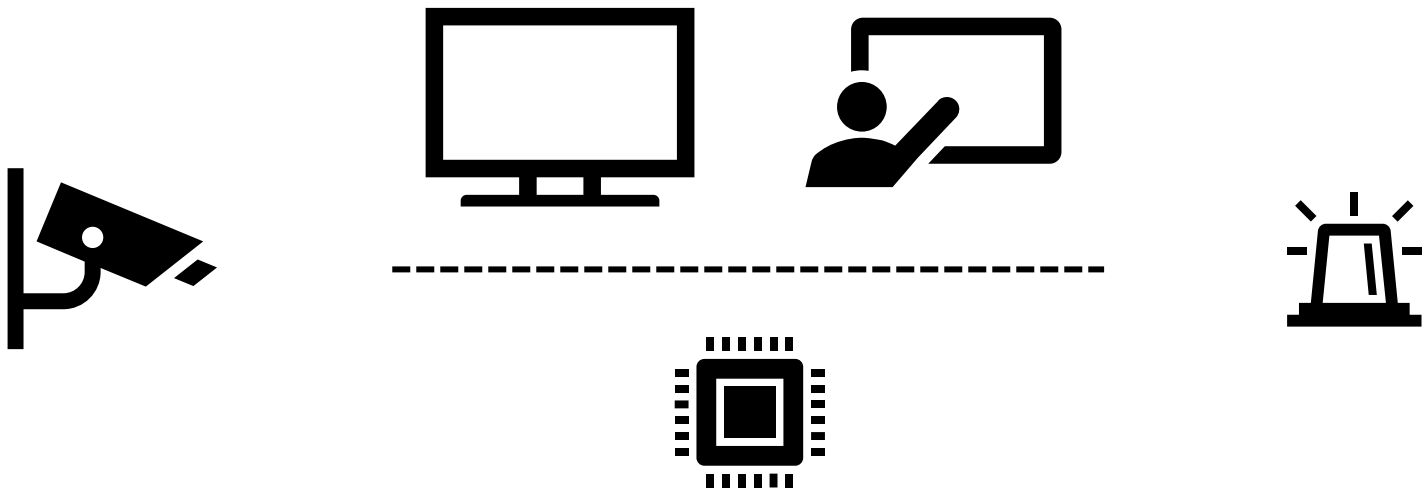
- **What is Video Anomaly Detection?**
  - ✓ VAD is to detect the anomalous activities such as crimes, fighting, road accidents, riots as well as the anomalous entities such as weapons at sensitive places, and abandoned objects
  - ✓ VAD is playing an increasingly important role in intelligent surveillance systems to reduce the manual work of liv e monitoring



Zhu, Yi, and Shawn Newsam. "Motion-Aware Feature for Improved Video Anomaly Detection." arXiv preprint arXiv:1907.10211 (2019).

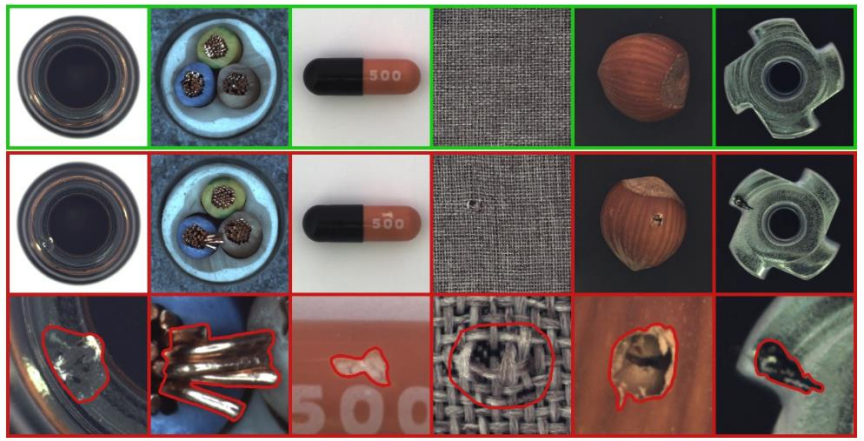
Video Anomaly Detection

- **Necessity of Video Anomaly Detection?**
  - ✓ The total number of surveillance cameras in Seoul is 144,513, and the trend continues to increase.
  - ✓ VAD is important for automated abnormal behavior detection because it is very difficult for monitoring staff to c heck for abnormal behavior while watching all CCTV.

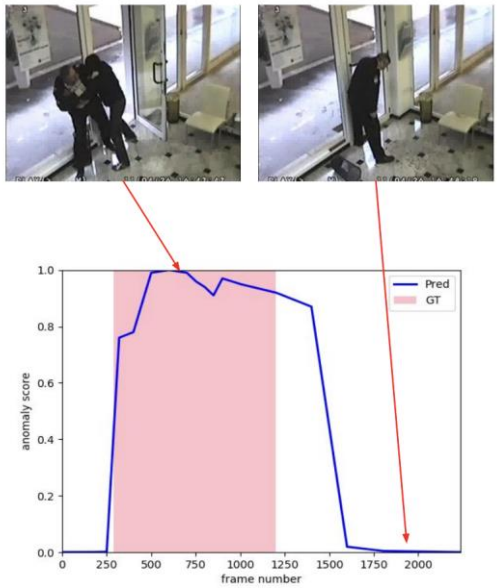


Video Anomaly Detection

- **What is the difference between image and video?**
  - ✓ Consists of continuous frames (images) -> Temporal information
  - ✓ The definition of abnormality varies greatly depending on the situation, behavior, etc.



VS



Fighting

Introduction

- **Datasets**
  - ✓ The initial anomaly datasets consisted of a small number of videos because it is extremely difficult or even impossible to collect sufficient and comprehensive abnormal data.
  - ✓ For every dataset, the definition of anomal events is diverse and ambiguous.

Dataset	# of Train	# of Test	Anomaly types	# of Scenes
UCSD Ped 1	6,800	7,200	5	1
UCSD Ped 2	2,550	2,010	5	1
CHUK Avenue	15,328	15,324	5	1
ShanghaiTech Campus	274,515	42,883	11	13
Ubnormal	116,087	92,640	22	29
UCF-Crime	12,631,211	1,110,182	13	

Introduction

- **UCSD Ped1, 2**
  - ✓ The UCSD Anomaly Detection Dataset was acquired with a stationary camera mounted at an elevation, overlooking p edestrian walkways.
  - ✓ Pedestroan 1 dataset includes 34 training videos and 36 testing videos with 40 irregular events.
  - ✓ Pedestrian 2 dataset contains 16 training videos and 12 testing videos with 12 abnormal events.
  - ✓ Commonly occurring anomalies include bikers, skaters, small carts, and people walking across a walkway or in the gr ass that surrounds it.



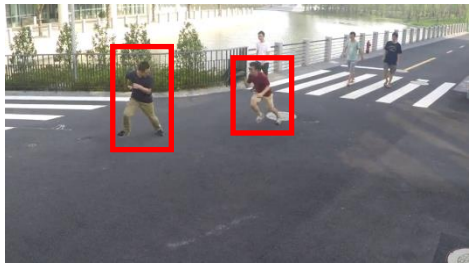

Introduction

- **CHUK Avenue**
  - ✓ Avenue Dataset contains 16 training videos and 21 testing videos with a total of 47 abnormal events, including throwing objects, bikers and running.



- **ShanghaiTech Campus**
  - ✓ ShanghaiTech Campus dataset has 13 scenes with complex light conditions and camera angles. It contains 330 training videos and 107 testing ones with 130 abnormal events.

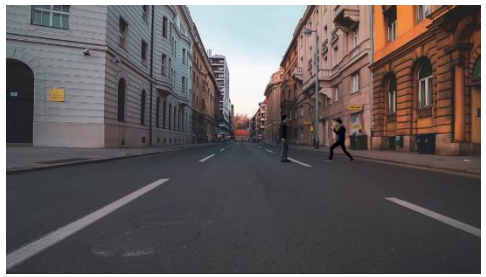




Introduction

- **UBnormal**
  - ✓ The UBnormal benchmark is generated using the Cinema4D software, comprising a new dataset of 29 virtual scenes with 236, 902 video frames.
  - ✓ The example of anomaly events include running, falling, fighting, sleeping, crawling, having a seizure, dancing, stealing etc.

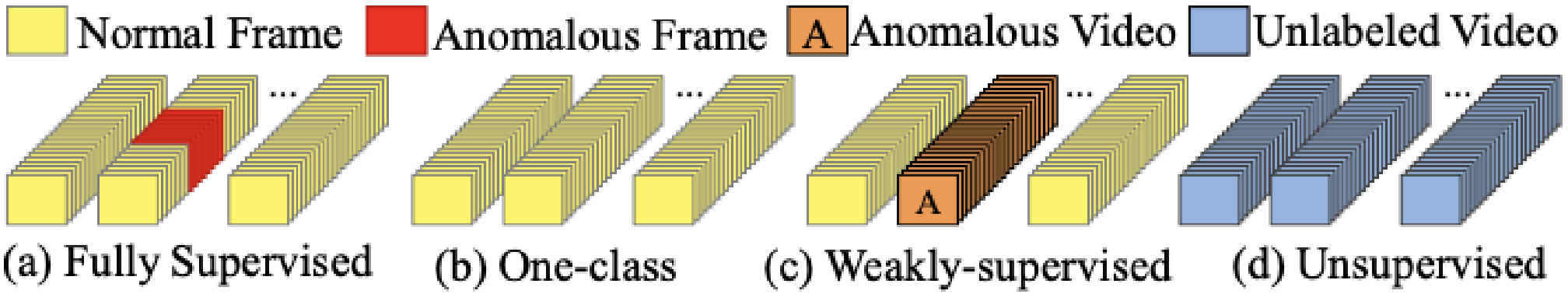




- **UCF-Crime**
  - ✓ UCF-Crime dataset consists of long untrimmed surveillance videos which cover 13 realworld anomalies, including Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism.

Introduction

- **Four Learning Method**
  - ✓ Fully Supervised: frame-level normal/abnormal annotations in the training data
  - ✓ One-class: only normal training data
  - ✓ Weakly-supervised: video-level normal/abnormal annotations
  - ✓ Unsupervised:no training data annotations



Zaheer, M. Zaigham, et al. "Generative cooperative learning for unsupervised video anomaly detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Learning Temporal Regularity in Video Sequences (CVPR, 2016)

- **Method**
  - ✓ One-class (Reconstruction based)
- **Issue**
  - ✓ Learning temporal visual characteristics of meaningful or salient moments is very challenging as the definition of such moments is ill-defined i.e., visually unbounded
- **Solution**
  - ✓ Authors address this problem by modeling temporal regularity of videos with limited supervision, rather than modeling the sp arse irregular or meaningful moments in a supervised manner.

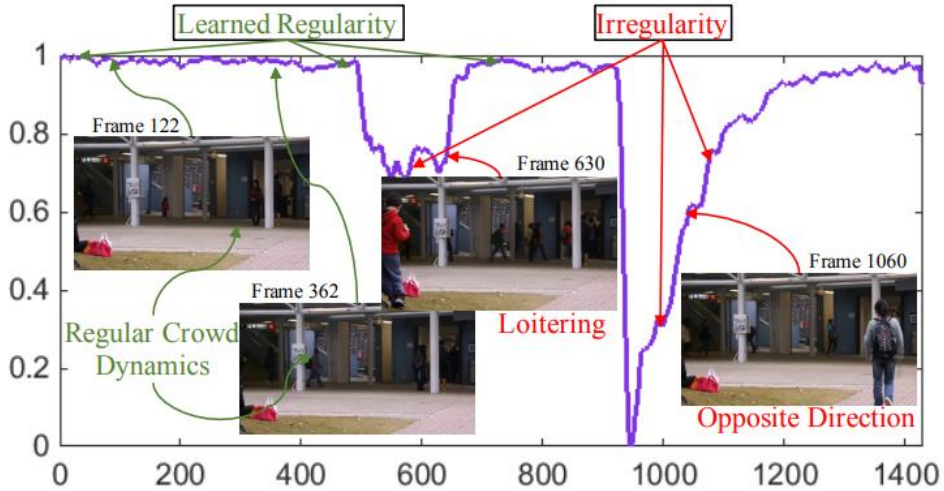


Figure 1. Learned regularity of a video sequence. Y-axis refers to regularity score and X-axis refers to frame number. When there are irregular motions, the regularity score drops significantly (from CUHK-Avenue dataset [8]).

Learning Temporal Regularity in Video Sequences (CVPR, 2016)

- **Method**
  - ✓ Authors show that an autoencoder effectively learns the regular dynamics in long-duration videos and can be applied to identify irregularity in the videos.
  - ✓ The learned autoencoder reconstructs regular motion with low error but incurs higher reconstruction error for irregular motions.
  - ✓ The reconstruction error is used to measure the regularity score that can be further analyzed for different applications.

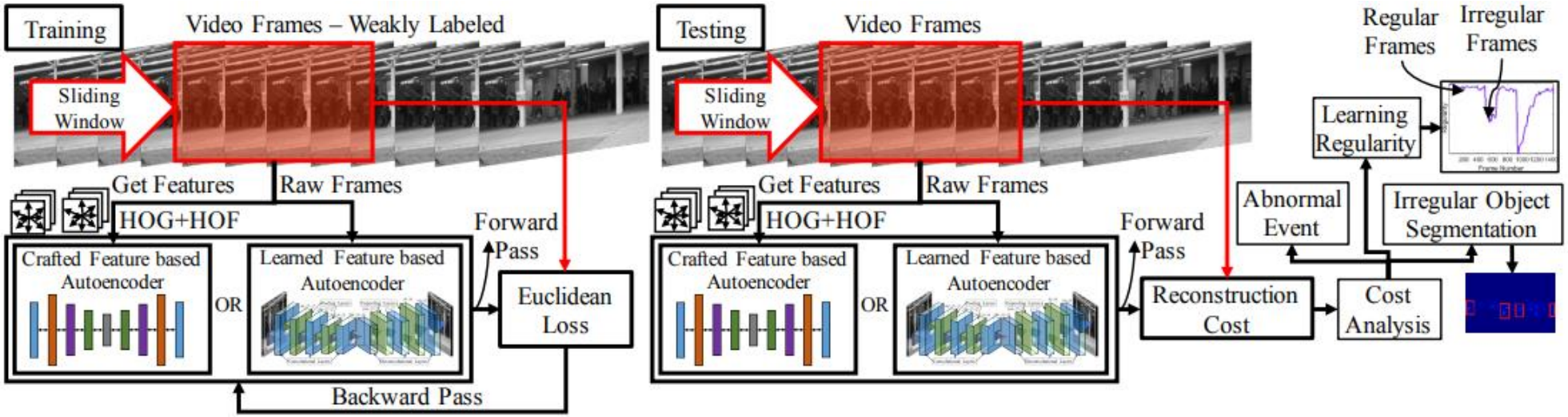


Figure 2. Overview of our approach. It utilizes either state-of-the-art motion features or learned features combined with autoencoder to reconstruct the scene. The reconstruction error is used to measure the regularity score that can be further analyzed for different applications.

Future Frame Prediction for Anomaly Detection - A New Baseline (CVPR, 2018)

- **Method**
  - ✓ One-class (Prediction based)
- **Issue**
  - ✓ Reconstruction-based methods usually learn a deep neural network with an Auto-Encoder way and they enforce it to reconstruct normal events with small reconstruction errors. But the capacity of deep neural network is high, and larger reconstruction errors for abnormal events do not necessarily happen.
- **Solution**
  - ✓ Author introduce a future video frame prediction based on its historical observation for video anomaly detection

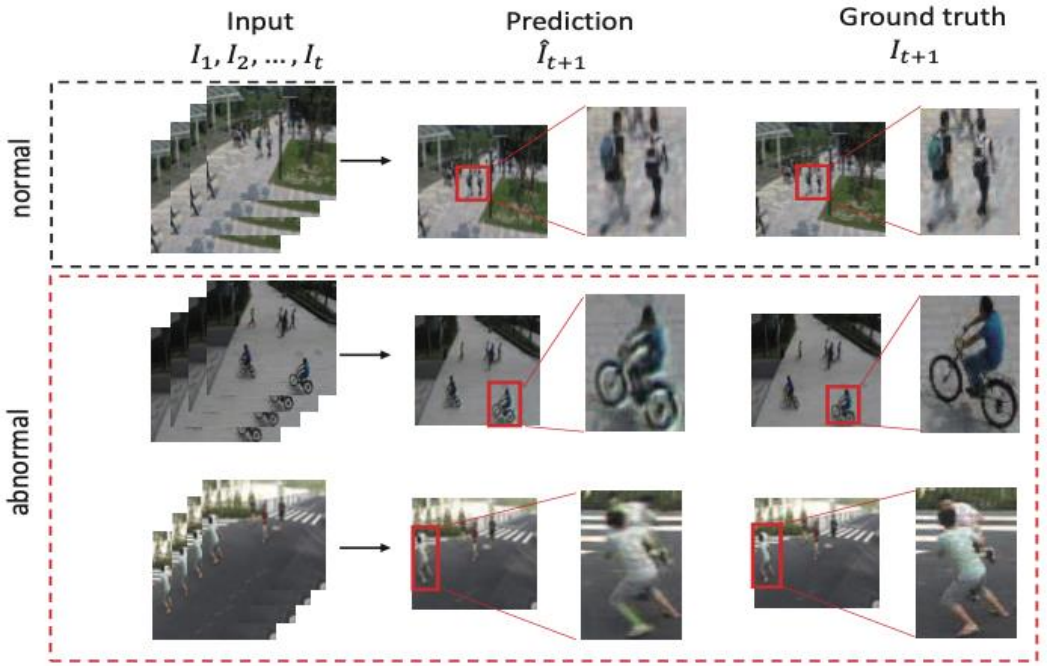
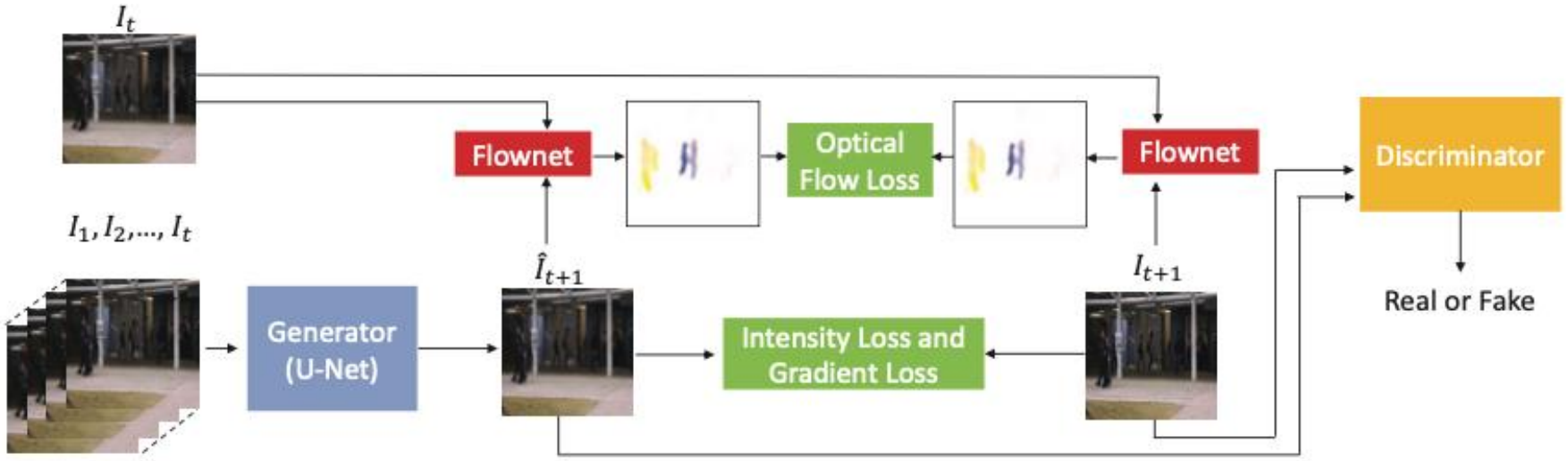


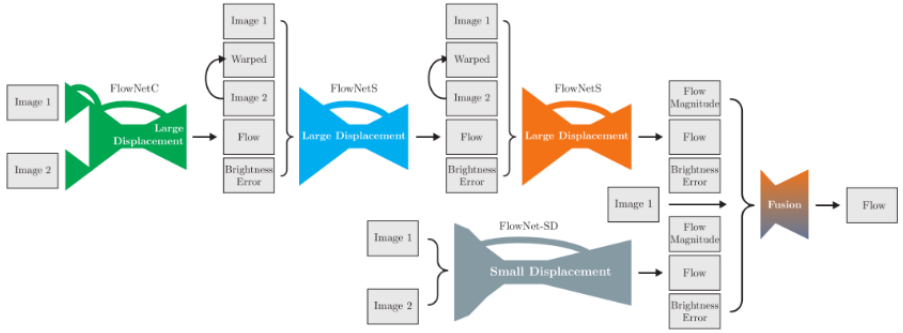
Figure 1. Some predicted frames and their ground truth in normal and abnormal events. Here the region is walking zone. When pedestrians are walking in the area, the frames can be well predicted. While for some abnormal events (a bicycle intrudes/ two men are fighting), the predictions are blurred and with color distortion. Best viewed in color.

Future Frame Prediction for Anomaly Detection - A New Baseline (CVPR, 2018)

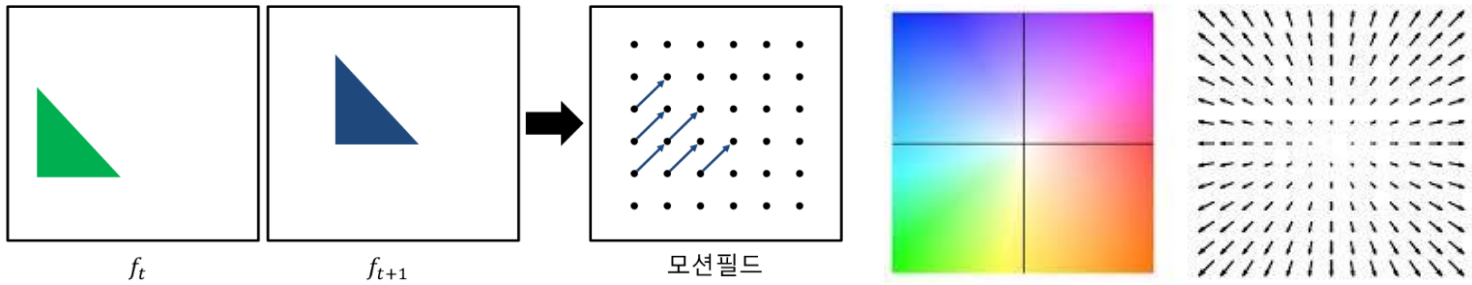
- **Architecture**
  - ✓ FFP is consisted of U-Net, Flownet, Discriminator.
  - ✓ Input is t frames stacked, ouput is a t+1 future frame
  - ✓ U-Net: A predictor that can well predict the future frame for normal training data
  - ✓ Discriminator: discriminate real or fake frame to generate more successful future frame
  - ✓ Flownet: pretrained network to estimate optical flow that means pixel by pixel motion

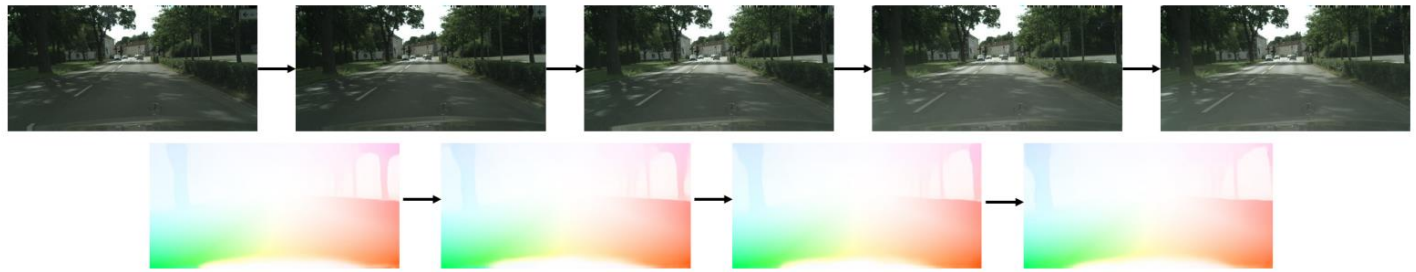


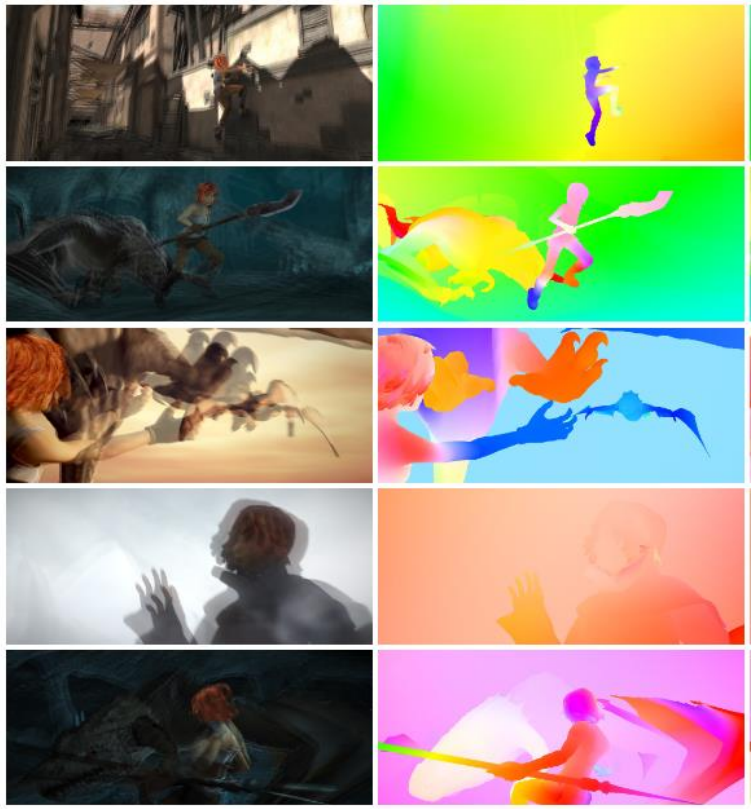
Future Frame Prediction for Anomaly Detection



- **FlowNet**
  - ✓ A deep learning model that generates FlowMap, an image that represents the change in the x-direction and the change in the y-direction between the two frames per pixel.



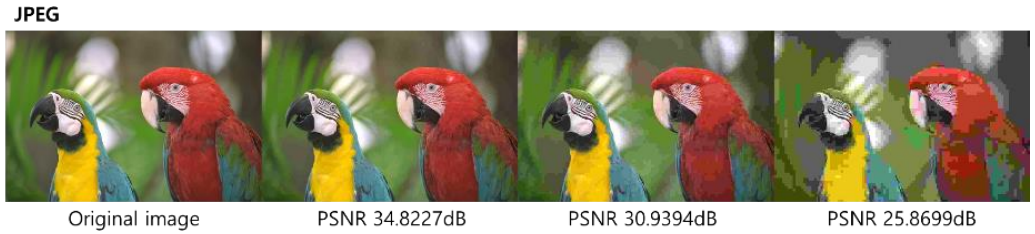


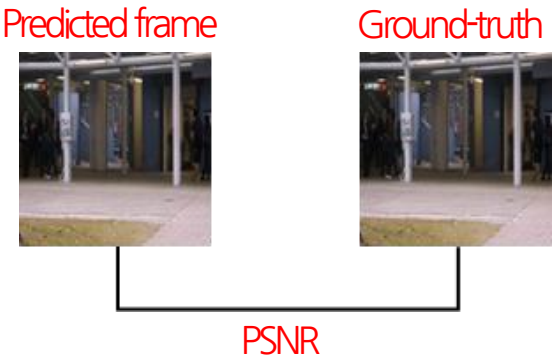


Future Frame Prediction for Anomaly Detection - A New Baseline(CVPR, 2018)

- **Testing**
  - ✓ Peak Signal-to-noise ratio (PSNR)
    - represents the quality of an image by normalizing Mean Squared Error (MSE) to pixel information scale; higher values indicate better quality.

$$PSNR(I, \hat{I}) = 10 \log_{10} \frac{[\max_{\hat{I}}]^2}{\frac{1}{N} \sum_{i=0}^{N} (I_i - \hat{I}_i)^2}$$





Real-world anomaly detection in surveillance videos (CVPR, 2018)

- **Method**
  - ✓ Weakly-supervised Method

- **Issue**
  - ✓ The assumption that all normal patterns are learned and th e patterns that deviate from them are considered abnorma l is not always valid.
  - ✓ The boundary between normal and abnormal is ambiguou s and the same behavior can be considered normal or abn ormal

- **Solution**
  - ✓ training data of normal and anomalous events with weakly -labeled video (video-level label)
  - ✓ Multiple instance learning

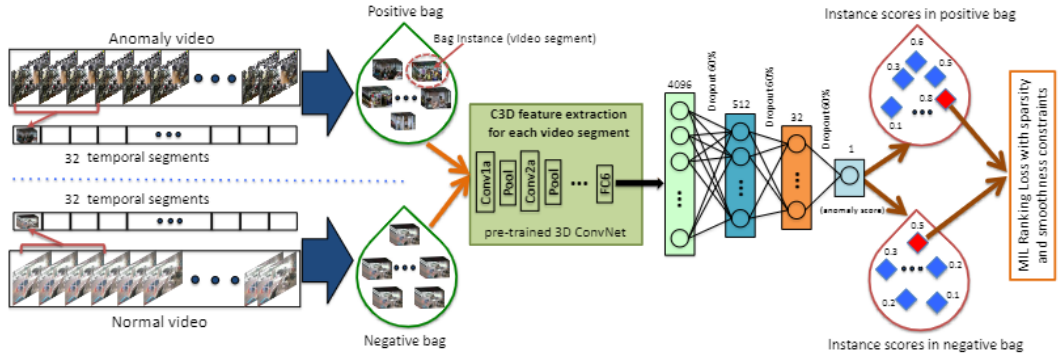


Figure 1. The flow diagram of the proposed anomaly detection approach. Given the positive (containing anomaly somewhere) and negative (containing no anomaly) videos, we divide each of them into multiple temporal video segments. Then, each video is represented as a bag and each temporal segment represents an instance in the bag. After extracting C3D features [36] for video segments, we train a fully connected neural network by utilizing a novel ranking loss function which computes the ranking loss between the highest scored instances (shown in red) in the positive bag and the negative bag.

Real-world anomaly detection in surveillance videos (CVPR, 2018)

- **Method**
  - ✓ Multiple Instance Learning
  - ✓ Using Convolution 3D

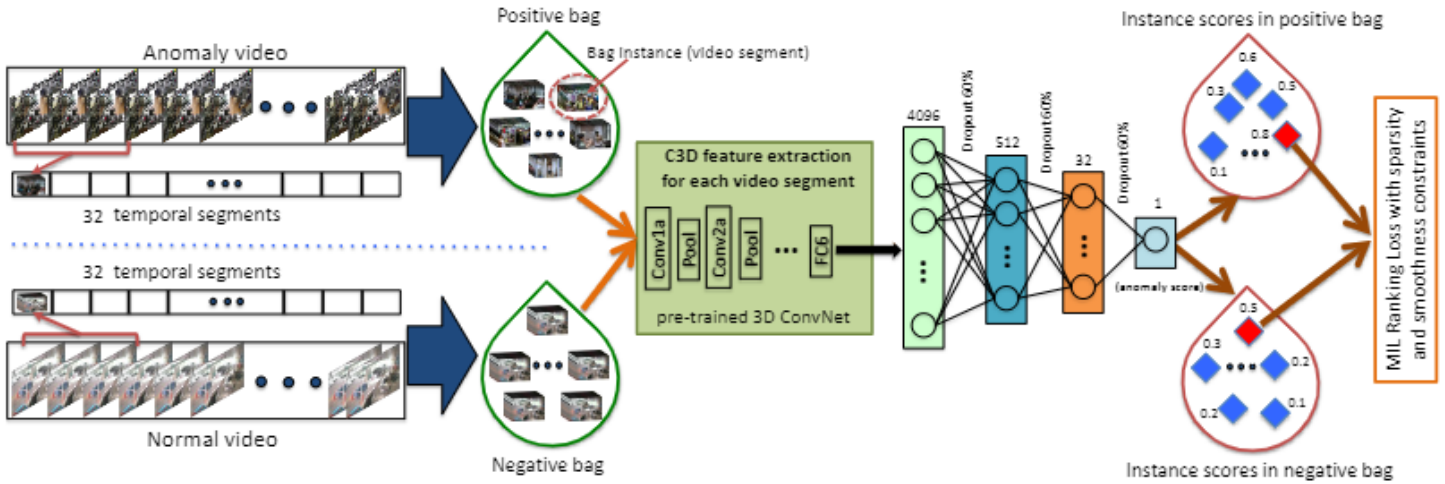


Figure 1. The flow diagram of the proposed anomaly detection approach. Given the positive (containing anomaly somewhere) and negative (containing no anomaly) videos, we divide each of them into multiple temporal video segments. Then, each video is represented as a bag and each temporal segment represents an instance in the bag. After extracting C3D features [36] for video segments, we train a fully connected neural network by utilizing a novel ranking loss function which computes the ranking loss between the highest scored instances (shown in red) in the positive bag and the negative bag.

Real-world anomaly detection in surveillance videos (CVPR, 2018)

- C3D

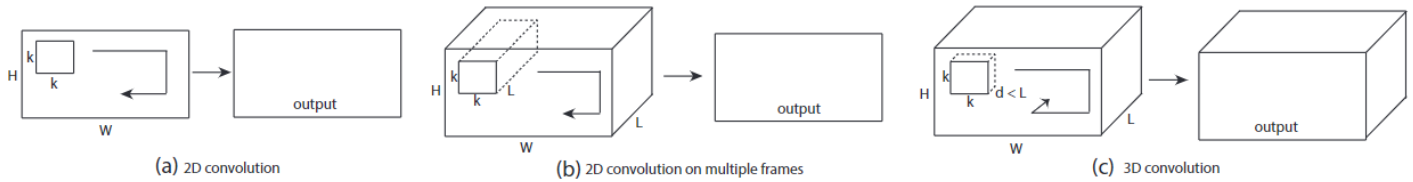
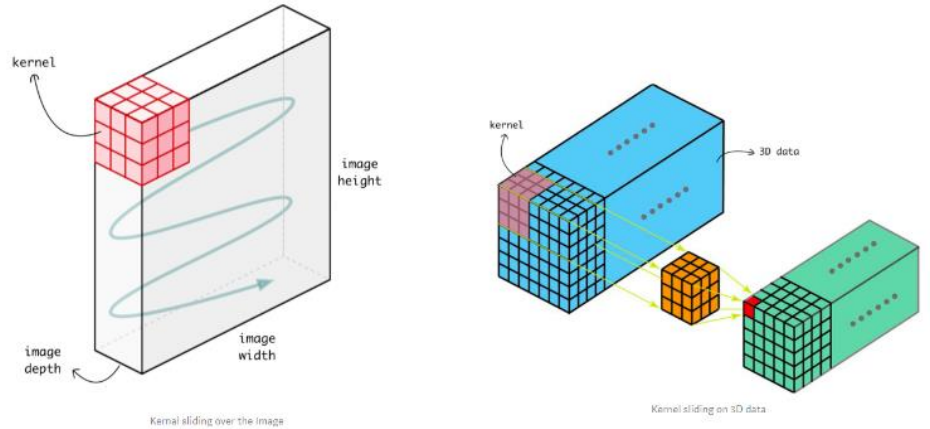


Figure 1. **2D and 3D convolution operations**. a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.



Real-world anomaly detection in surveillance videos (CVPR, 2018)

- Multiple Instance Learning

$$f(\mathcal{V}_a) > f(\mathcal{V}_n), \qquad \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) > \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i),$$

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i))$$

$$+\lambda_1 \overbrace{\sum_{i}^{(n-1)} (f(\mathcal{V}_a^i) - f(\mathcal{V}_a^{i+1}))^2}^{①} + \lambda_2 \overbrace{\sum_{i}^{n} f(\mathcal{V}_a^i)}^{②},$$

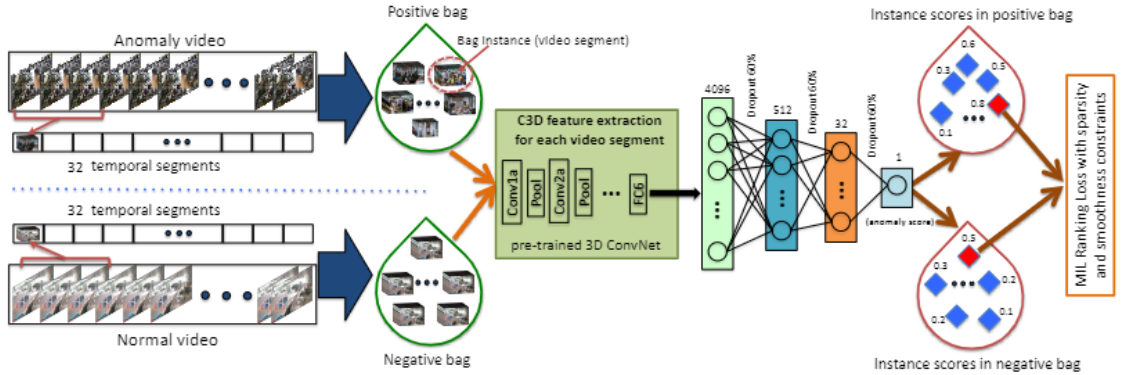


Figure 1. The flow diagram of the proposed anomaly detection approach. Given the positive (containing anomaly somewhere) and negative (containing no anomaly) videos, we divide each of them into multiple temporal video segments. Then, each video is represented as a bag and each temporal segment represents an instance in the bag. After extracting C3D features [36] for video segments, we train a fully connected neural network by utilizing a novel ranking loss function which computes the ranking loss between the highest scored instances (shown in red) in the positive bag and the negative bag.

Generative Cooperative Learning for Unsupervised Video Anomaly Detection(CVPR, 2022)

- **Method**
  - ✓ Unsupervised method
- **Issue**
  - ✓ One-class classification methods are usually unsuitable for complex problems with diverse multiple classes and a wide range of dynamic situations often found in vide o surveillance.
  - ✓ Weakly supervised methods require video-level label, w hich are relatively cost-effective, yet remain impractical in many real-world applications.
- **Solution**
  - ✓ The authors propose an unsupervised video anomaly detection method that does not require labeled traini ng data.

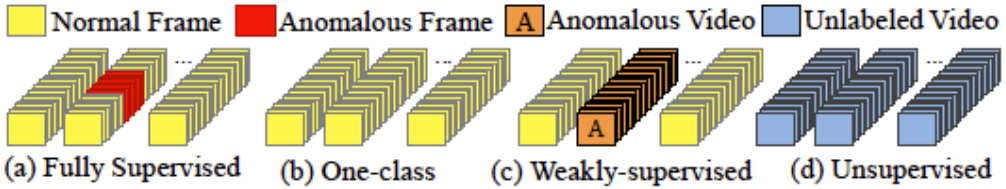
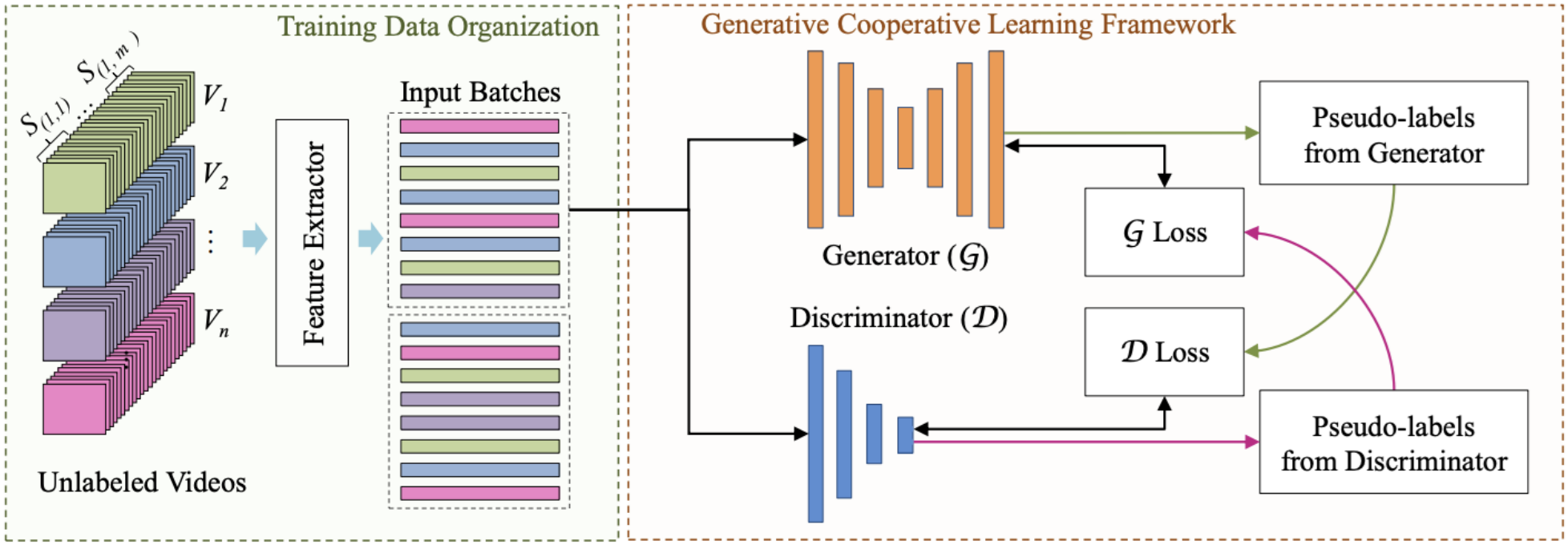


Figure 1. Different training modes for video anomaly detection: (a) Fully supervised mode requires frame-level normal/abnormal annotations in the training data. (b) One-Class Classification (OCC) requires only normal training data. (c) Weakly supervised mode requires video-level normal/abnormal annotations. (d) Un-supervised mode requires no training data annotations.

Generative Cooperative Learning for Unsupervised Video Anomaly Detection(CVPR, 2022)

- **Architecture**
  - ✓ Generative Cooperative Learning approach for Anomaly Detection (GCL) comprises a feature extractor, a gen erator network, a discriminator network, and two pseudo-label generators, which essentially get trained in a mutually cooperative manner.
  - ✓ Output is snippet-level anomaly score ranging from 0 to 1 generated by the discriminator.

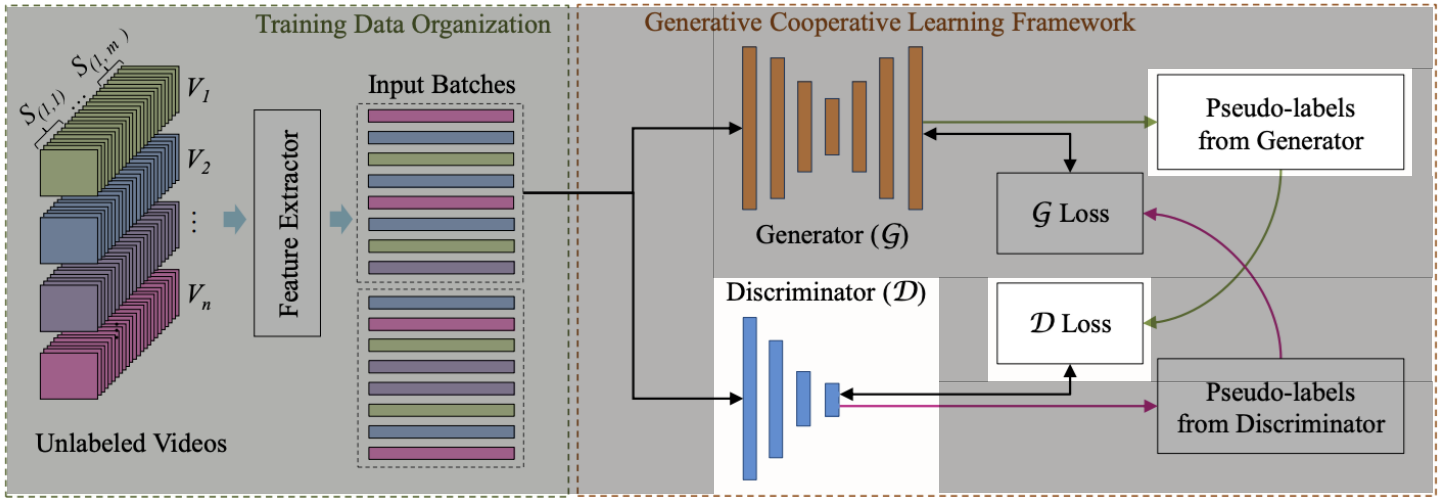


Generative Cooperative Learning for Unsupervised Video Anomaly Detection(CVPR, 2022)

- **Discriminator**
  - ✓ The binary classification network used as the discriminator is trained using the pseudo annotations from generator
  - ✓ Model's output is snippet-level anomaly score that range 0 to 1.
- **Pseudo labels from generator**
  - ✓ In proposed collaborative learning, pusedo labels from generator are created to train discriminator.

$$\mathcal{L}_r = \frac{1}{b}\sum_{q=1}^{b}\mathcal{L}_G^q,\ \mathcal{L}_G^q = ||f_{i,j}^q - \widehat{f}_{i,j}^q||_2, \qquad l_G^q = \begin{cases} 1, & \text{if } \mathcal{L}_G^q \geq \mathcal{L}_G^{th} \\ 0, & \text{otherwise .} \end{cases}$$

- **Loss**
  - ✓ Loss is binary cross entropy. $\mathcal{L}_D = \frac{-1}{b}\sum_{q=1}^{b} l_G^q \ln \widehat{l}_{i,j}^q + (1 - l_G^q)\ln(1 - \widehat{l}_{i,j}^q),$



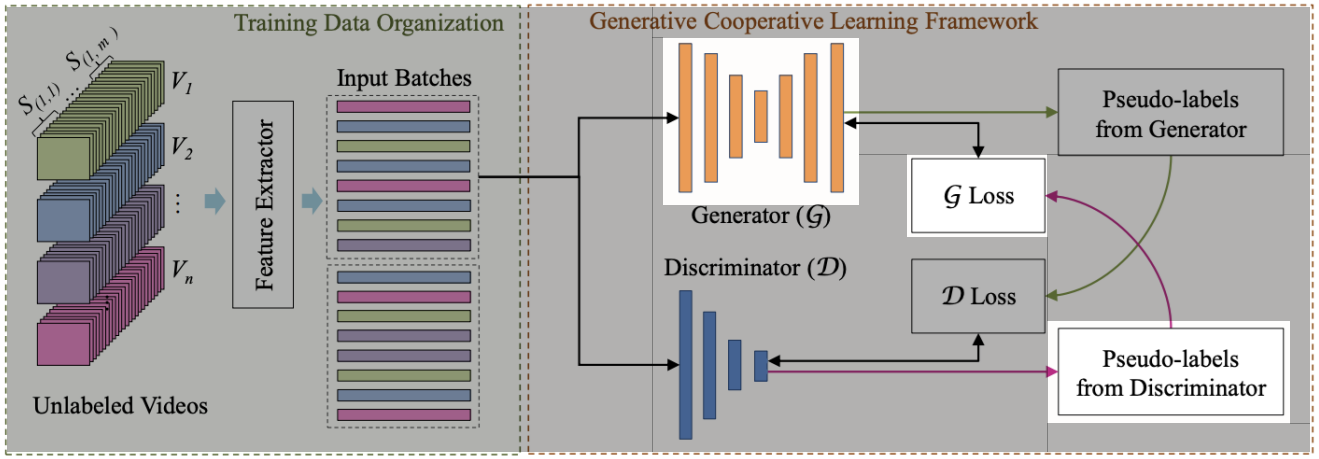
Generative Cooperative Learning for Unsupervised Video Anomaly Detection(CVPR, 2022)

- **Generator**
  - ✓ Generator based autoencoder takes features as input and produces reconstructions of those features as output.
- **Pseudo labels from discriminator**
  - ✓ Pseudo labels from discriminator are used to improve the reconstruction discrimination capability of generator.

$$l_D^q = \begin{cases} 1, & \text{if } \widehat{p}_{i,j}^q \geq \mathcal{L}_D^{th} \\ 0, & \text{otherwise}, \end{cases} \qquad t_{i,j}^q = \begin{cases} f_{i,j}^q, & \text{if } l_D^q = 0 \\ \mathbf{1} \in \mathbb{R}^d, & \text{if } l_D^q = 1, \end{cases}$$

- **Loss**
  - ✓ Loss is binary cross entropy.

$$\mathcal{L}_G = \frac{1}{b}\sum_{q=1}^{b} ||t_{i,j}^q - \widehat{f}_{i,j}^q||_2,$$



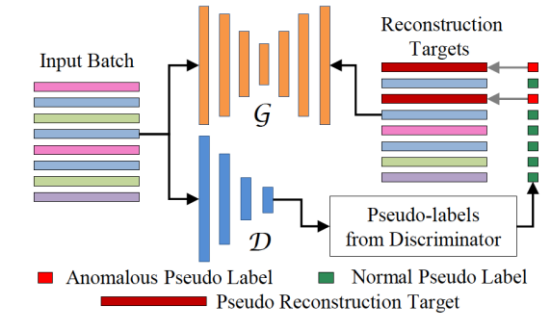
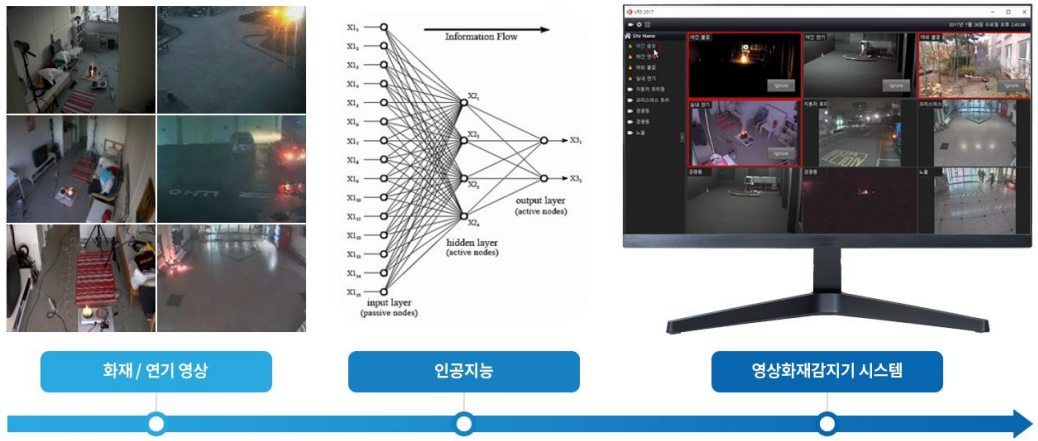
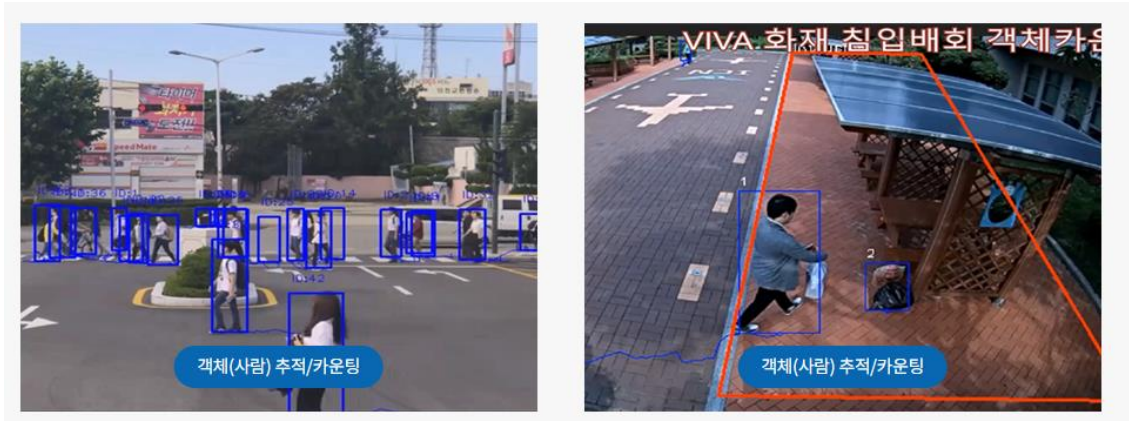


Figure 3. Negative learning in GCL: $\mathcal{G}$ is constrained to not learn the reconstruction of anomalies using Pseudo Reconstruction Targets (PRT). Based on the pseudo-labels produced by $\mathcal{D}$, PRT are generated for the anomalous inputs while normal targets are used for the normal inputs to guide the training of $\mathcal{G}$.

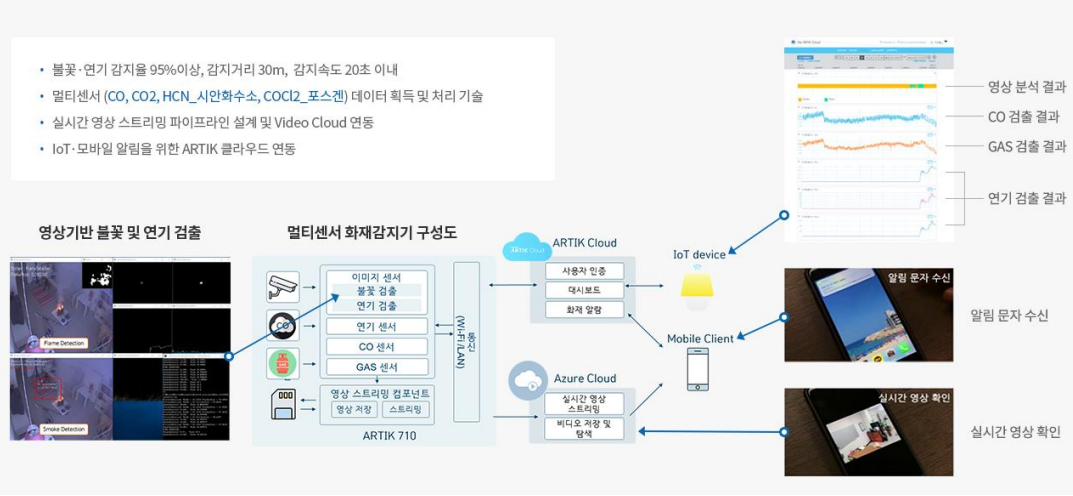
Industry

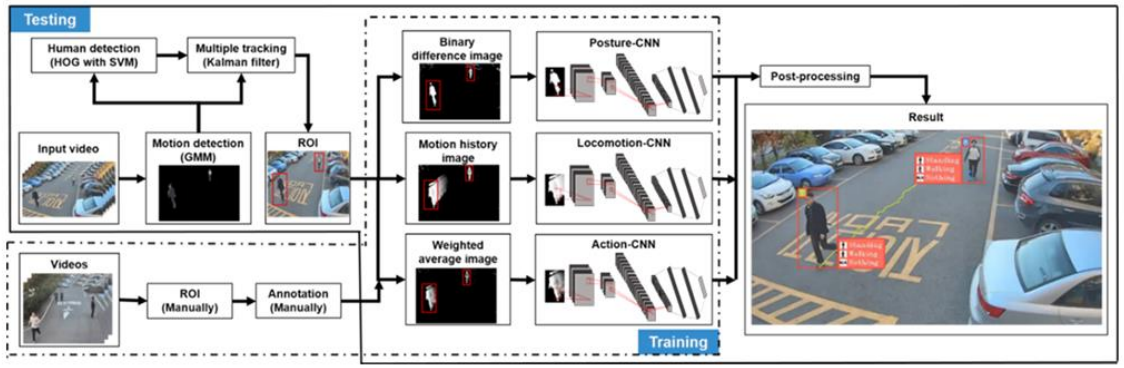
- VISIONIN





- 불꽃·연기 감지율 95%이상, 감지거리 30m, 감지속도 20초 이내
- 멀티센서 (CO, CO2, HCN_시안화수소, COCl2_포스겐) 데이터 획득 및 처리 기술
- 실시간 영상 스트리밍 파이프라인 설계 및 Video Cloud 연동
- IoT·모바일 알림을 위한 ARTIK 클라우드 연동





Overall scheme of real-time action detection model

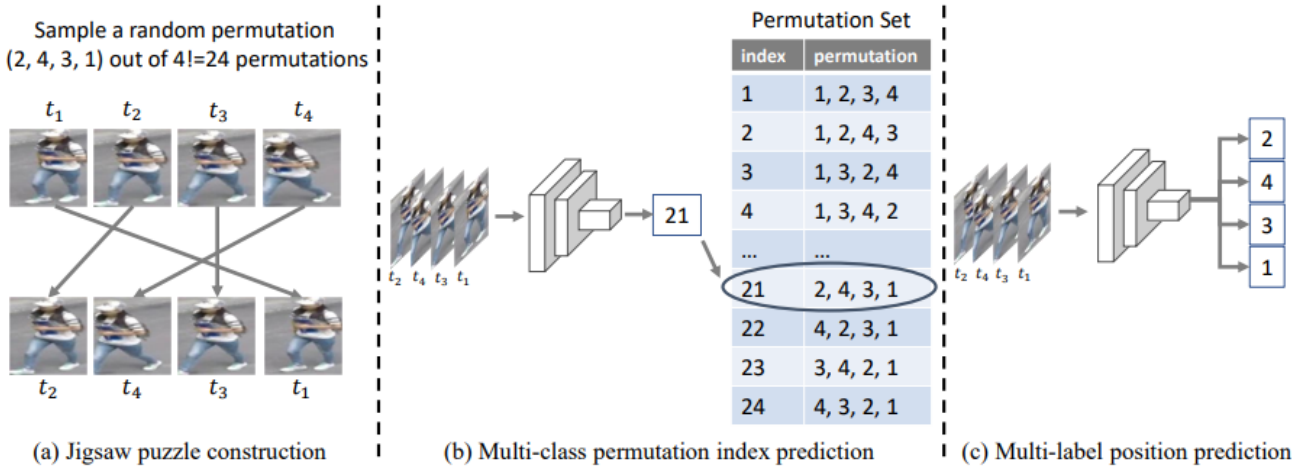
Trends

Application In Database Systems
(CSI8782.01-01)

Data Engineering Lab
Multi Modal Deep Learning Team

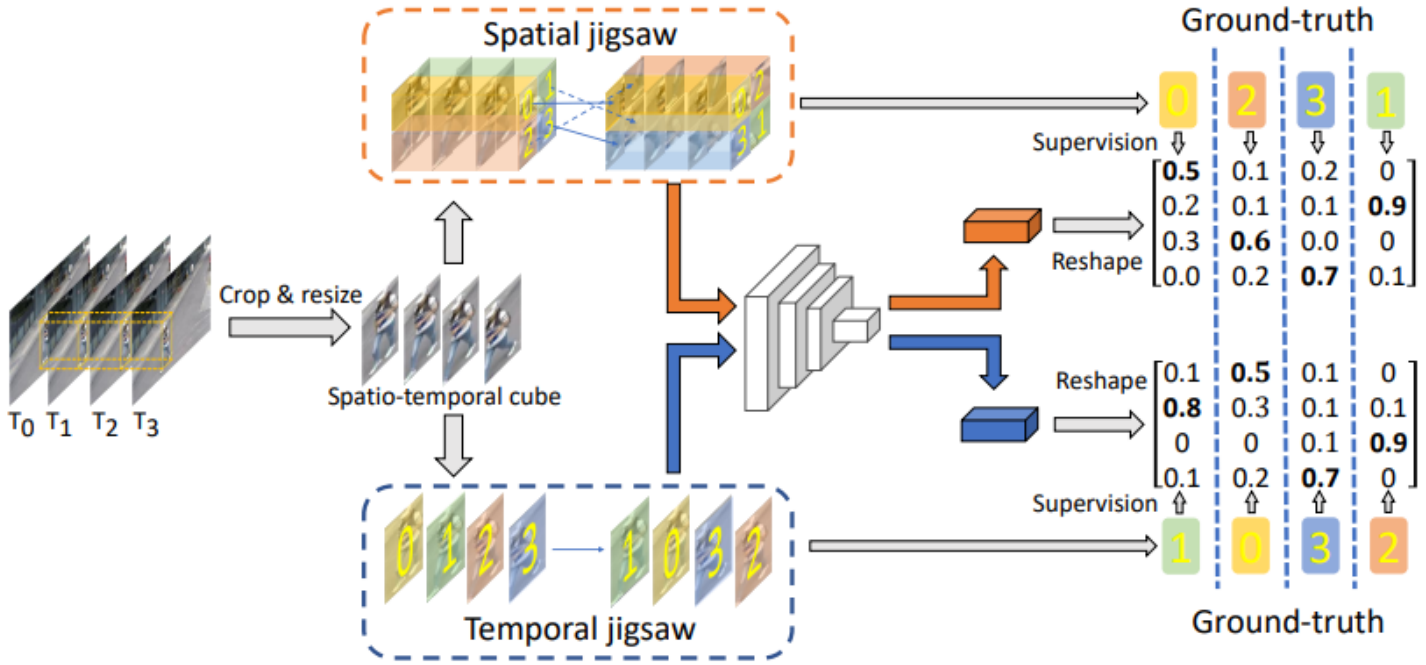
Video Anomaly Detection by Solving Decoupled Spatio-Temporal Jigsaw Puzzles (ECCV, 2022)

- **Method**
  - ✓ One-class (Self-supervised learning)
- **Issue**
  - ✓ Reconstruction and prediction based models are only trained to match normal examples, their inherent generalization abilities still make the anomalies well reconstructed or predicted.
  - ✓ Existing SSL methods are easy to solve, preventing the network from learning highly discriminative representations for VAD.
- **Solution**
  - ✓ This paper addresses VAD by solving an intuitive yet challenging pretext task, i.e., spatio-temporal jigsaw p uzzles, which is cast as a multi-label fine-grained clas sification problem.



Video Anomaly Detection by Solving Decoupled Spatio-Temporal Jigsaw Puzzles (ECCV, 2022)

- **Architecture**
  - ✓ The method contains three steps: object-centric cube extraction, puzzle construction, and puzzle solving.
  - ✓ For each cube, Authors apply spatial or temporal shuffling to construct the corresponding spatial or temporal jigsaw puzzle.
  - ✓ Finally, a convolutional neural network, acting as a jigsaw solver, attempts to recover the original sequence from its spatially or temporally permuted version.
  - ✓ The proposed method is equivalent to solving a multi-label classification problem and is trained in an end-to-end manner.



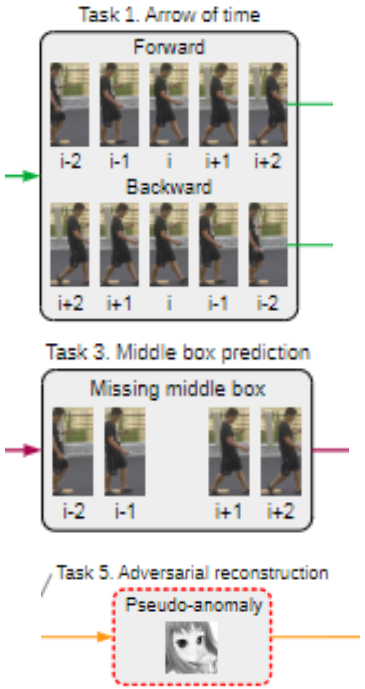
Video Anomaly Detection by Solving Decoupled Spatio-Temporal Jigsaw Puzzles (ECCV, 2022)

- **Ablation Study**
  - ✓ Authors can observe a trend of performance improvement when they increase l and n^2 in a certain range. However, when they increase l and n^2 further, the performance deteriorates.
  - ✓ Right table gives evidence that a proper design of the pretext task enabling fine-grained discrimination is essential for VAD. Authors set a more challenging task which requires the model to perceive every patch within a frame and every frame within a clip.

Exp. ID	l	n^2	Avenue	STC
B1	5	4	89.7	79.3
B2	7	4	90.2	80.2
B3	7	9	**92.2**	83.2
B4	9	4	88.6	81.1
B5	9	9	89.2	**84.3**
B6	9	16	87.9	80.4

Exp. ID	Spatial	Temporal	STC
D1	Rotation	Arrow of time	72.9
D2	Rotation	Temporal order verification	74.8
D3	Translation	Arrow of time	73.0
D4	Translation	Temporal order verification	75.6
D5	Translation	Jigsaw (ours)	81.1
D6	Jigsaw (ours)	Temporal order verification	78.3
D7	Jigsaw (ours)	Jigsaw (ours)	**84.3**

Table 3: Results of different numbers of frames/patches on STC and Avenue in terms of AUROC (%). l and n^2 denote the number of frames in an object-centric cube and the number of patches in the frames, respectively.



Feature Prediction Diffusion Model for Video Anomaly Detection (ICCV, 2023)

- **Method**
  - ✓ One-class (Prediction based)
- **Issue**
  - ✓ The GAN/AE-based methods suffer from the weak generative capacity, leading to more noise from the low-quality generated image, which reduces the performance.
  - ✓ Current SOTA methods heavily rely on foreground object or action information extracted by auxiliary models.
- **Solution**
  - ✓ The authors proposed diffusion model-based approach does not have this reliance and can make accurate feature prediction.

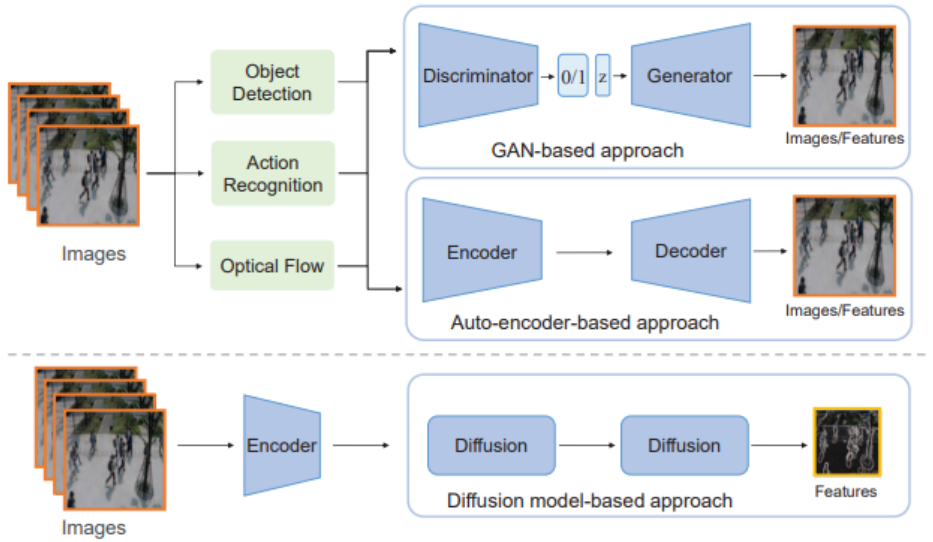
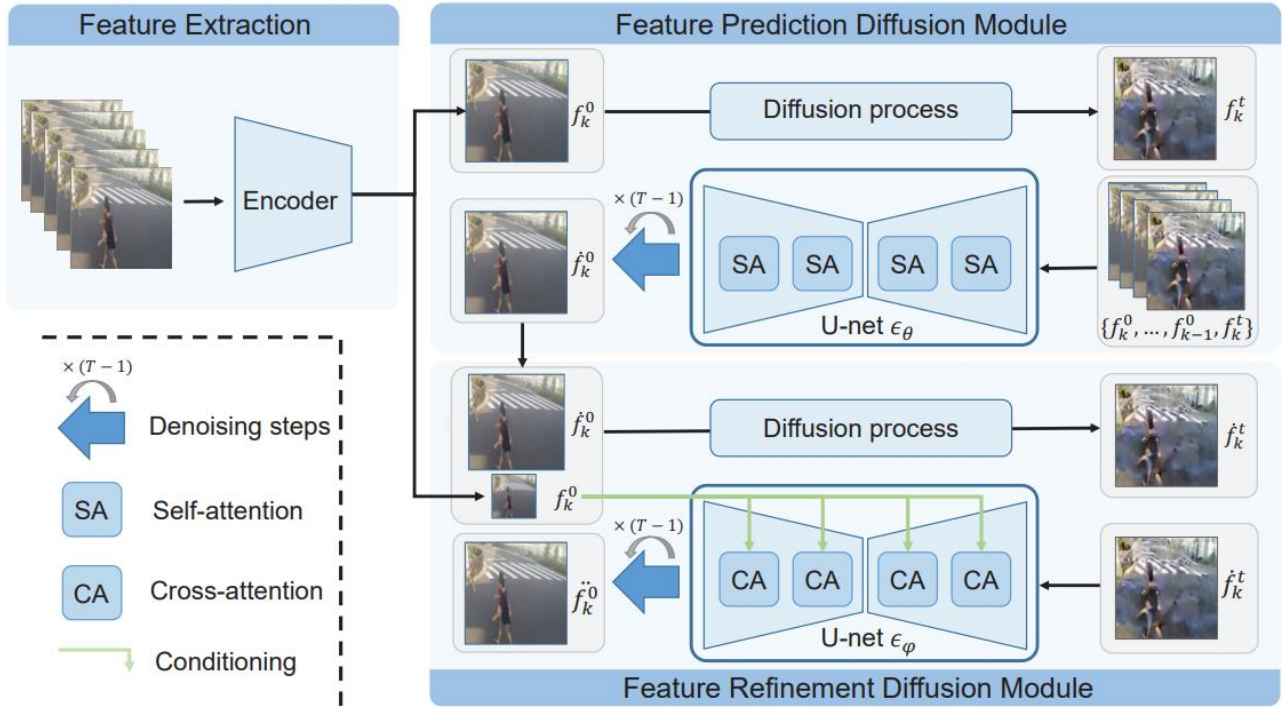


Figure 1. Overview of three generative VAD approaches. Existing state-of-the-art GAN-based or auto-encoder-based approaches heavily rely on foreground object or action information extracted by auxiliary models, such as object detection, action recognition or optical flow network, to generate features/images for effective performance. By contrast, our proposed diffusion model-based approach does not have this reliance and can make accurate feature prediction using only simple networks as encoder to extract basic image features as input.

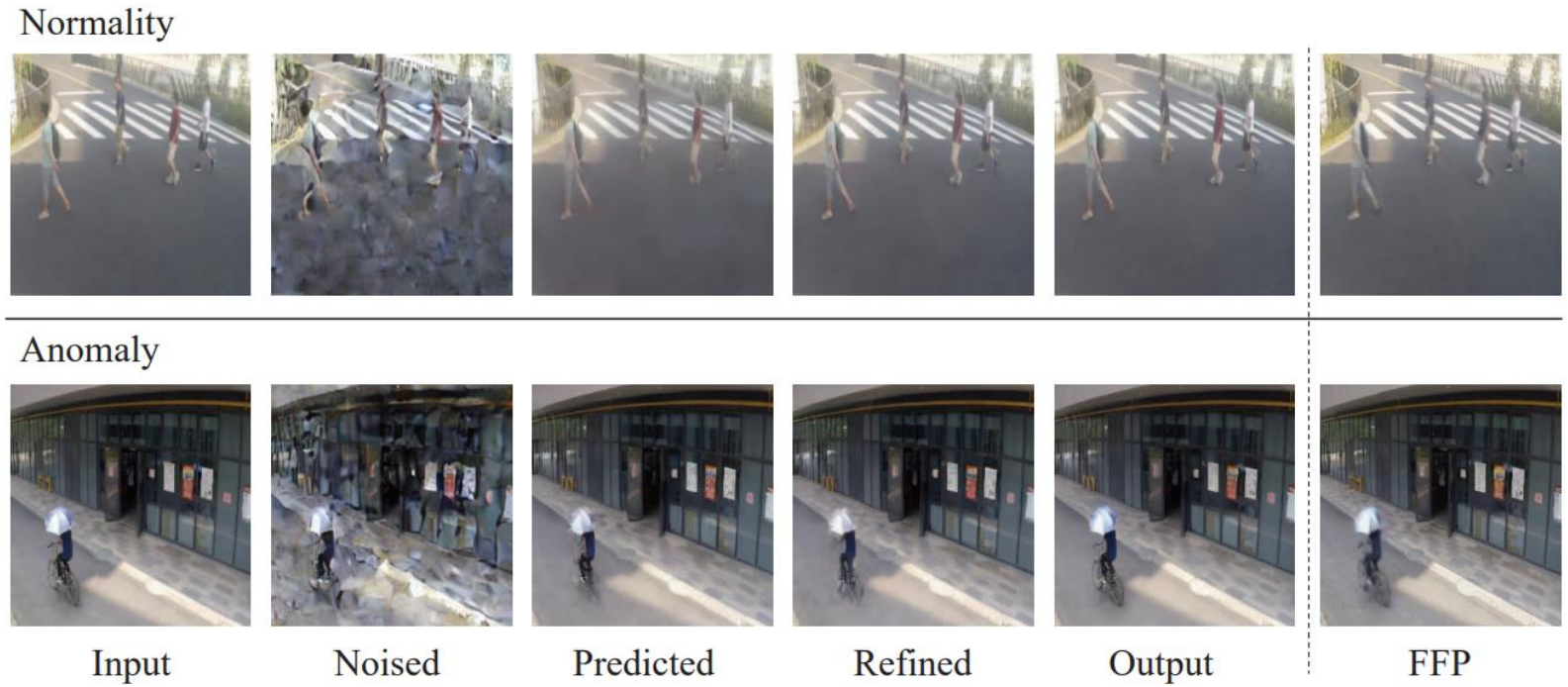
Feature Prediction Diffusion Model for Video Anomaly Detection (ICCV, 2023)

- **Architecture**
  - ✓ The feature prediction diffusion module adopts consecutive k features as input, in which only the last one accept the diffus ion process. With the temporal information from the consecutive frames, this module emphasizes learning the distribution of normal motion.
  - ✓ The feature refinement diffusion module takes the sampling output of previous module as input and the k-th original featu res as a condition for training, which focuses on the appearance learning.



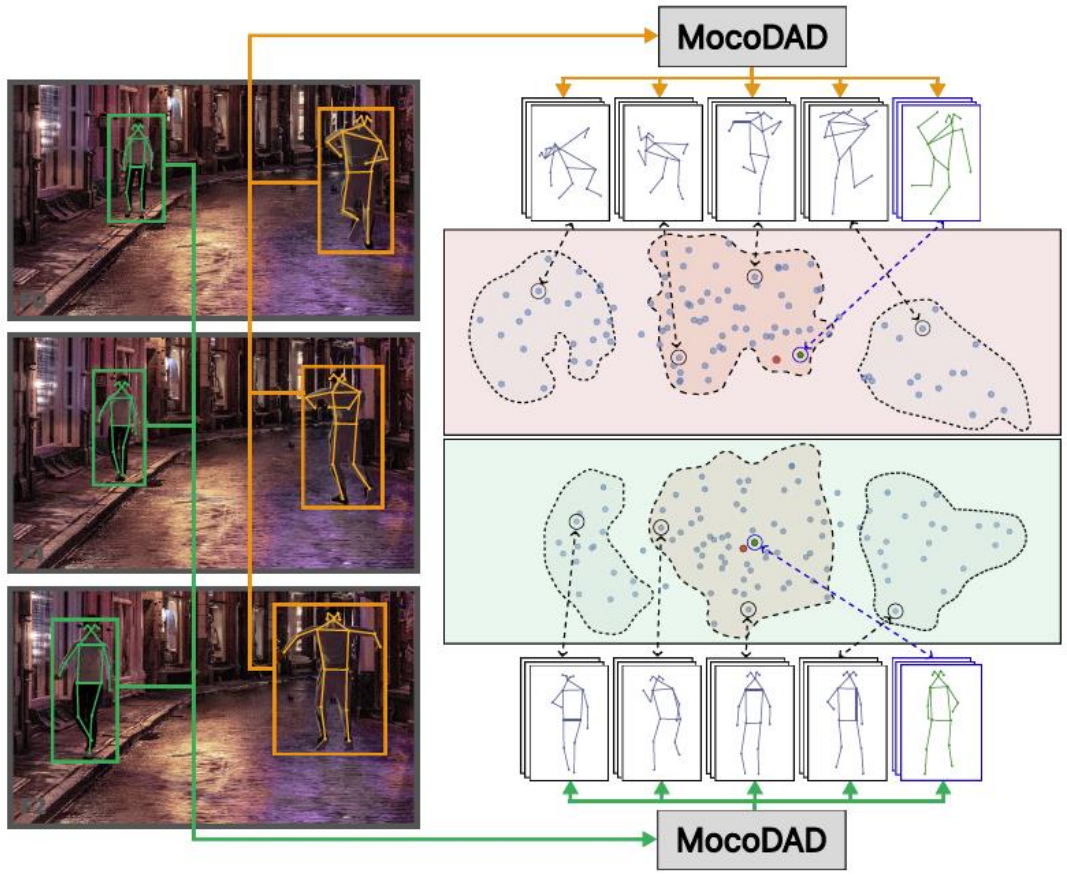
Feature Prediction Diffusion Model for Video Anomaly Detection (ICCV, 2023)

- **Visualization results**
  - ✓ **(Normality)** we can see that the predicted features well predict the position of each foreground object, but missing some detail information, such as legs and head of the person. These missing details are restored by the refinement module, as s hown in the refined column.
  - ✓ **(Anomaly)** The predicted feature in the third column also captures the global information of the foreground person but missing some details around this person. With the help of Refined module, the details of the bicycle, e.g., the two wheels are still blurred



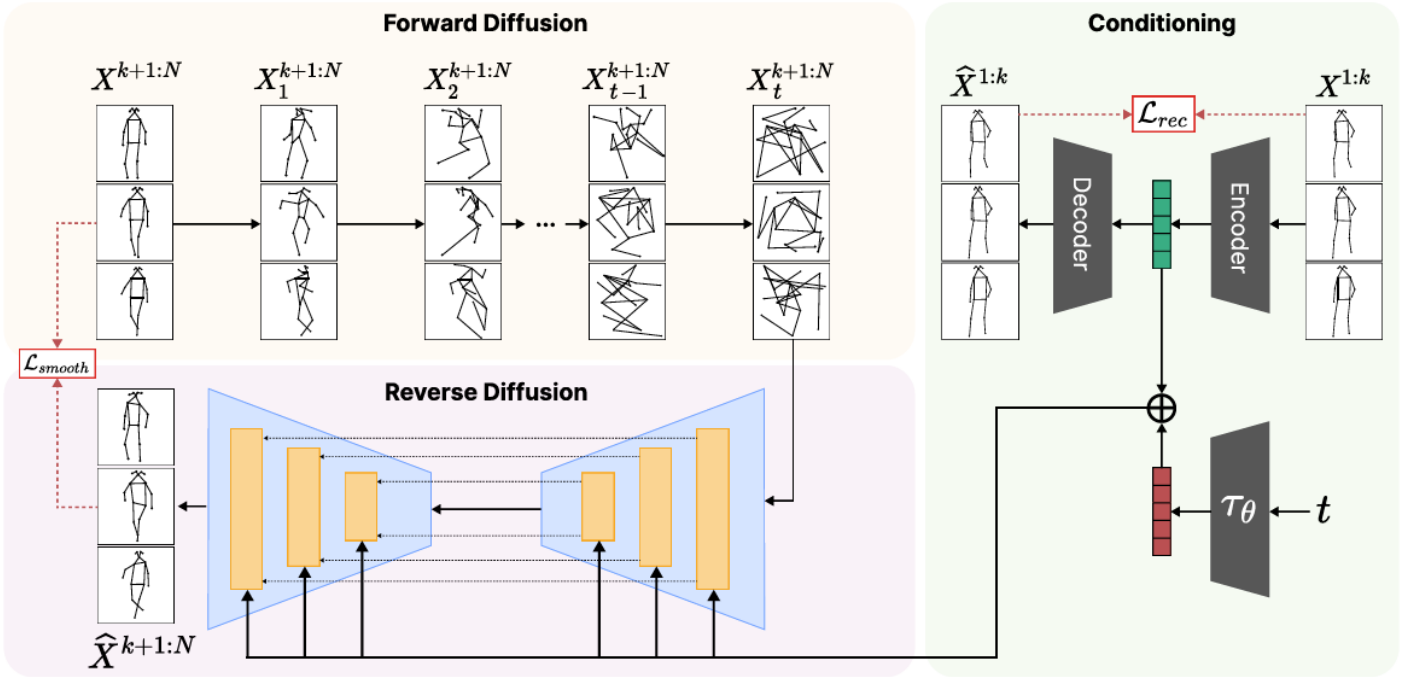
Multimodal Motion Conditioned Diffusion Model (CVPR, 2023)

- **Method**
  - ✓ One-Class (Prediction based)
- **Issue**
  - ✓ An "ideal" model for anomaly detection should consider that there are infinitely many anomalous and non-anom alous ways of performing an action.
  - ✓ Current sota OCC techniques fail to address this issue.
- **Solution**
  - ✓ Author proposed diffusion-based VAD model, called Mo CoDAD, which generates diverse multimodal motions.
  - ✓ Detect anomalies by comparing generated diverse motio ns with the original motion.



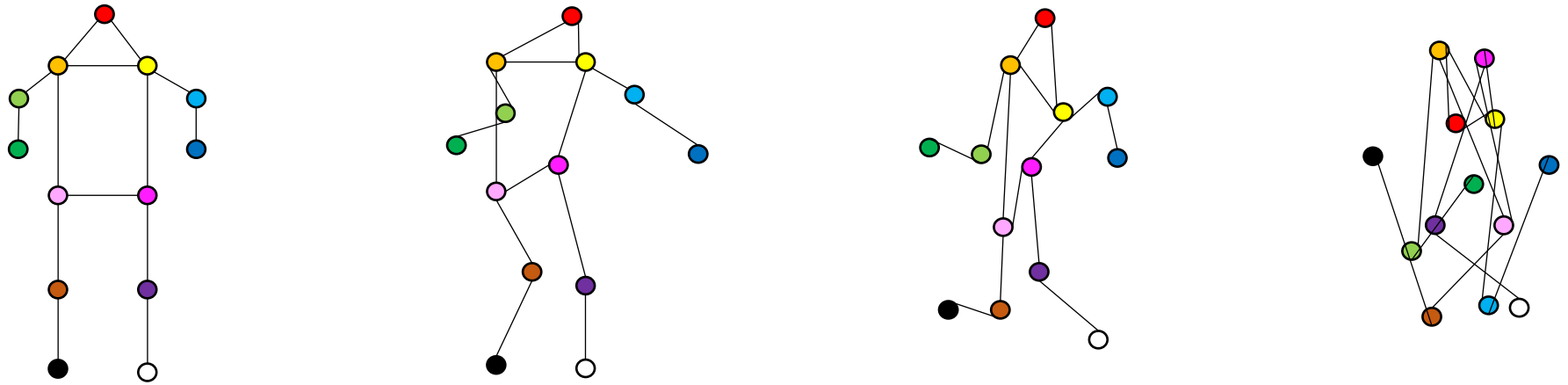
Multimodal Motion Conditioned Diffusion Model (CVPR, 2023)

- **Architecture**
  - ✓ Given a motion sequence, be it normal or anomalous, the sequence is split and the later (future) frames are corrupte d to become random noise.
  - ✓ MoCoDAD generates diverse future motion(skeleton architecture) sequences conditioned past motion sequence from gaussian noise vector.
  - ✓ MoCoDAD statistically aggregates them at inference to detect anomalies.



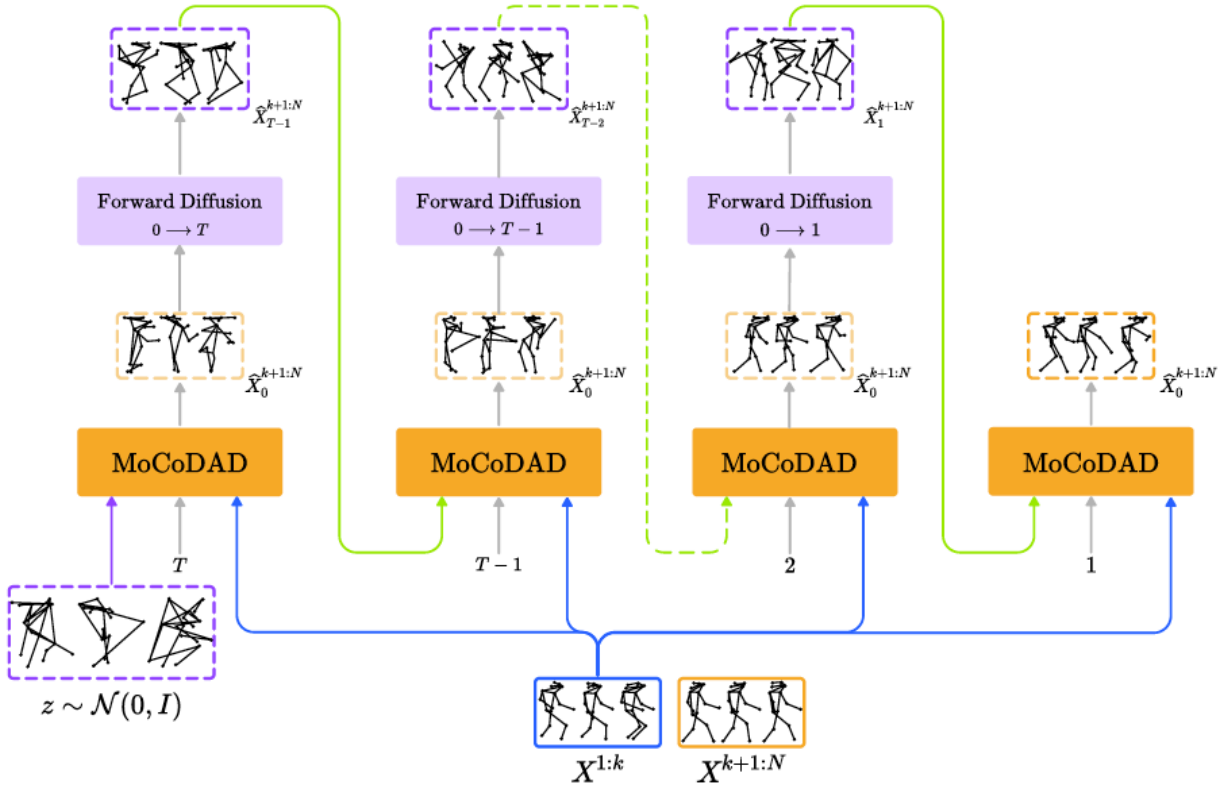
Multimodal Motion Conditioned Diffusion Model (CVPR, 2023)

- Add Noise



Multimodal Motion Conditioned Diffusion Model (CVPR, 2023)

- **Inference**
  - ✓ MoCoDAD denoises iteratively Gaussian noise vectors to predict diverse future motion sequences.
  - ✓ Compute the reconstruction errors between generated motions and the original motion.
  - ✓ The anomaly score aggregates the computed reconstruction errors.



TEVAD: Improved video anomaly detection with captions (CVPR, 2023)

- **Method**
  - ✓ Weakly-supervised method
- **Issue**
  - ✓ abnormal events in video are vaguely defined due to their ambiguous nature
  - ✓ previous methods do not consider the high-level sem antic meanings of the videos
- **Solution**
  - ✓ Proposed model generate the dense captions for snip pets and fed into pretrained text embedding network . These features are fused with the visual features to compute the anomaly scores

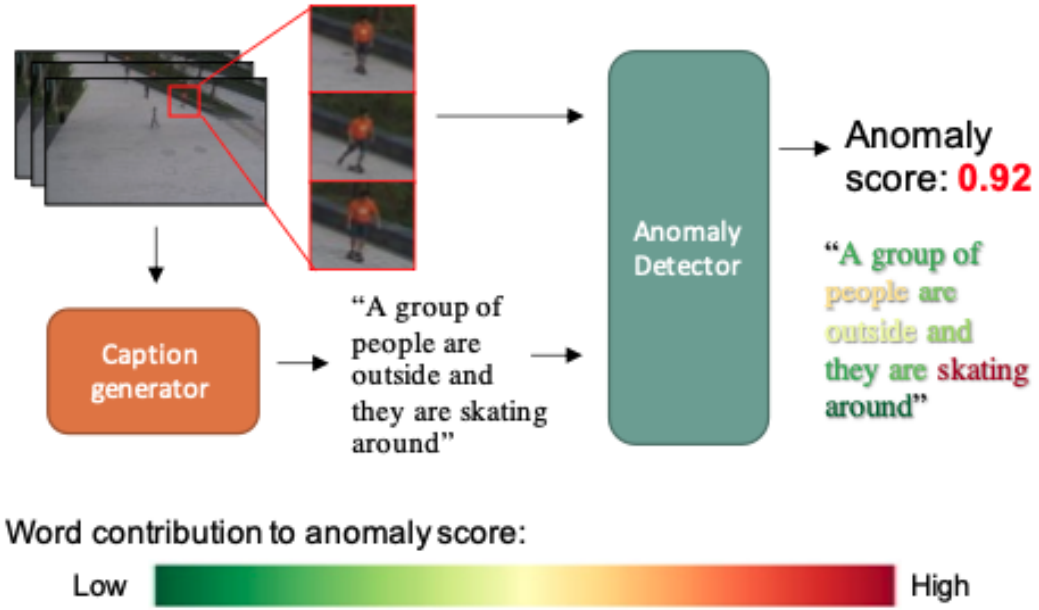
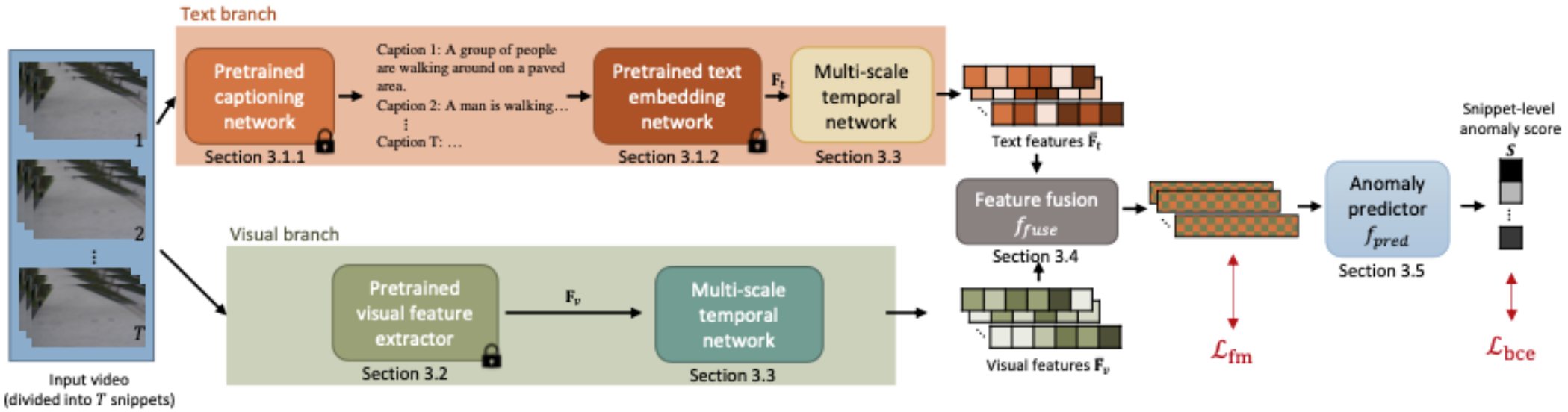


Figure 1. Our TEVAD first generates dense captions for snippets of a video, before using both visual and text modalities for video anomaly detection. The right side shows the predicted anomaly score and the contributions of each word to the prediction. The use of captions provides explainability to the model: the illustrated video is classified anomalous due to the "skating" action.

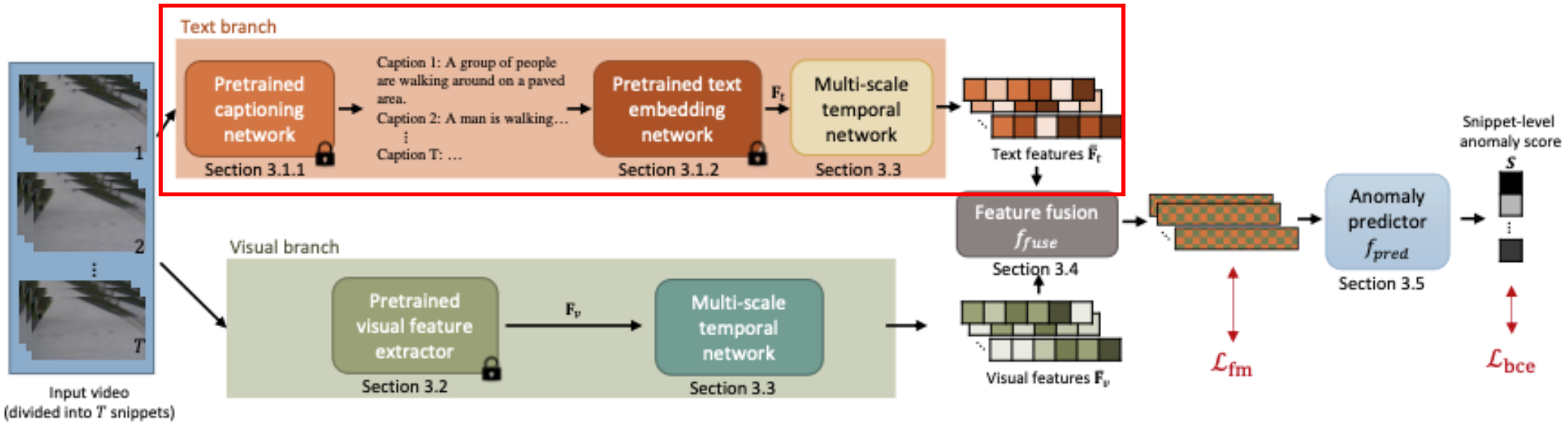
TEVAD: Improved video anomaly detection with captions (CVPR, 2023)

- **Architecture**
  - ✓ TEVAD is consisted of two separate branches and feature fusion, anomaly predictor
  - ✓ Given training video V, TEVAD splits each input video into T snippets to use input.
  - ✓ Output is snippet-level anomaly score that range 0 to 1



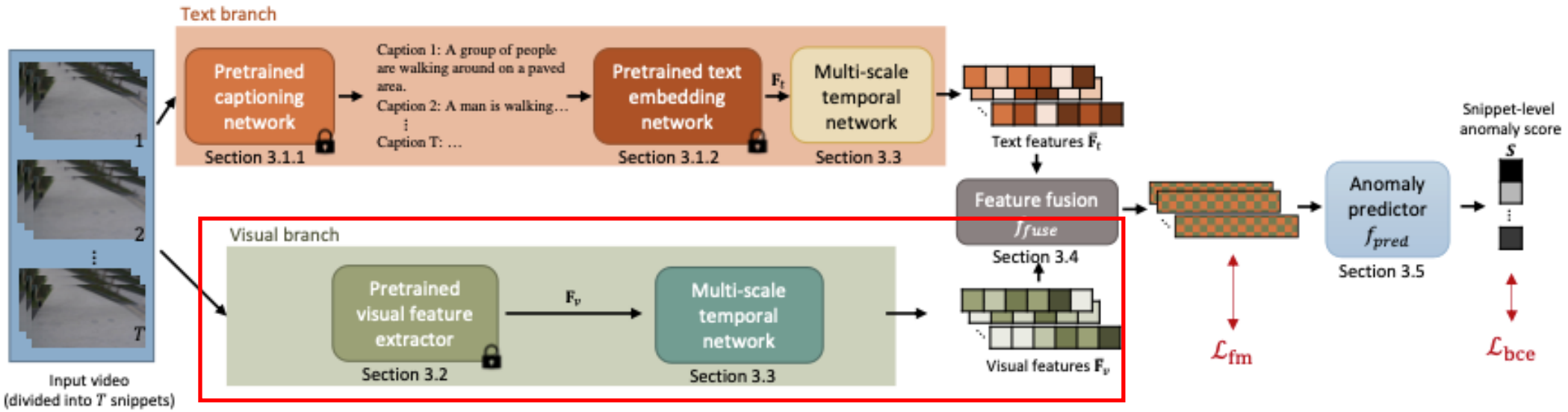
TEVAD: Improved video anomaly detection with captions (CVPR, 2023)

- **Text branch**
  - ✓ Text branch is consisted of video captioning model SwinBERT and text embedding model SimCSE and Mutl-scale temporal network(MTN)
  - ✓ SwinBERT generate dense captions for videos
  - ✓ SimCSE generate text feature vector from caption generated SwinBERT



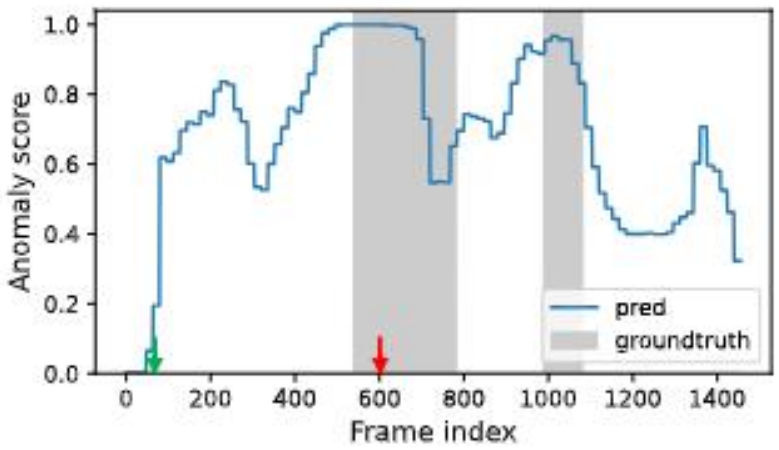
TEVAD: Improved video anomaly detection with captions (CVPR, 2023)

- **Visual branch**
  - ✓ Visual branch is consisted of I3D model and Multi-scale temporal network(MTN).
  - ✓ I3D is a commonly used visual feature extractor in VAD, pretrained on action recognition tasks.



TEVAD: Improved video anomaly detection with captions (CVPR, 2023)

- Result



"a man is walking in a store and he is drinking a beer bottle."

"a man is juggling a bottle of liquor and then he throws it to the camera."

0	0.02	0	-0.01	0.01
a	group	of	people	are
0.03	0.06	-0.01	0	-0.07
riding	bikes	around	a	walkway

(a)

0.01	0	0.01	0.01	0.01
a	man	is	holding	a
0.04	0.01	0.01	0.01	0
gun	and	then	he	shoots
0.01	0.01	0.01	0	
it	at	the	camera.	

(b)

0.04	0.03	0.03	0.02	0.02
a	group	of	people	are
0.03	0.03	0.02	0.04	0.04
in	a	room	and	one
0.03	0.03	0.09	0.03	
of	them	falls	down.	

(c)

Thank you