2025 동계 세미나

Video Anomaly Detection by memory and Jigsaw Puzzles

Pusan National University
Dept. of Electronic Engineering
Computer Vision & Signal Processing Lab

Presented By
이우진

CVSP Lab
Computer Vision & Signal Processing

Contents

- Paper review
  - Paper 1
    - Memorizing Normality to Detect Anomaly: Memory-augmented DeepAutoencoder for Unsupervised Anomaly Detection(MemAE)
  - Paper 2
    - Learning Memory-guided Normality for Anomaly Detection(MNAD)
  - Paper 3
    - Video Anomaly Detection by Solving DecoupledSpatio-Temporal Jigsaw Puzzles(STJP)

논문선정 이유

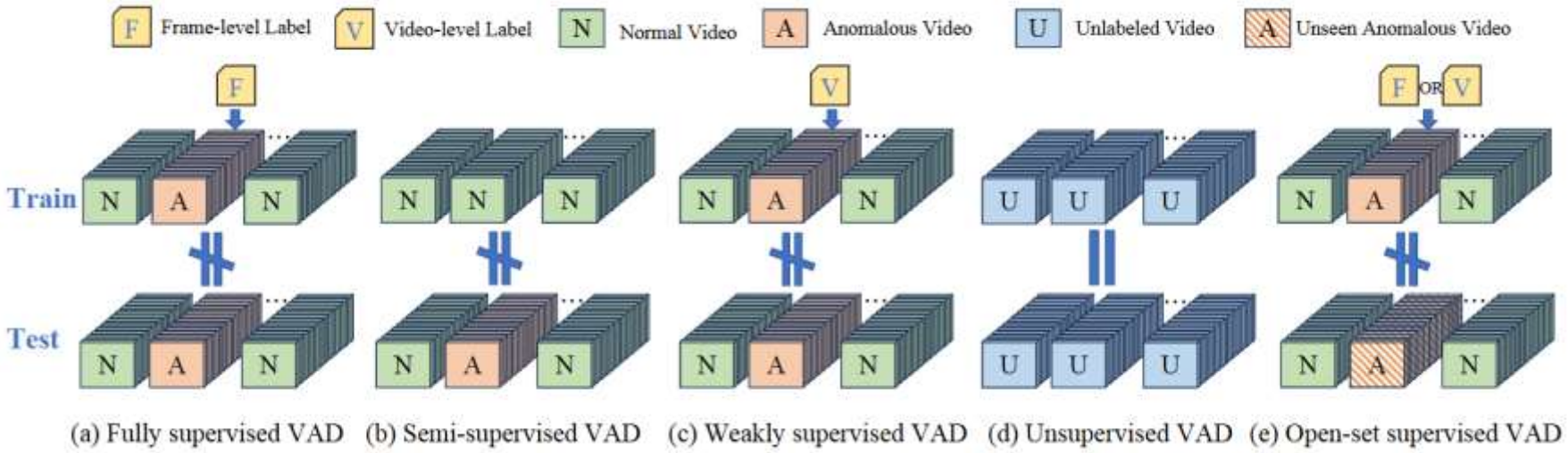
- MemAE: memory 기반의 VAD를 처음 제시한 논문이고, 해당분야에 지대한 영향을 끼쳤기 때문에 소개

- MNAD: MemAE의 한계점이라 할 수 있는 너무 많은 memory item을 획기적으로 줄이는 방식을 제안함으로써 memory 기반 vad를 효율적으로 개선한 논문이므로 선정

- STJP: 기존의 주류기법(reconstruction, prediction 등등)들을 완전히 벗어난 독창적인 방식으로 video anomaly detecting을 해낸 논문이라서 인상깊어서 소개

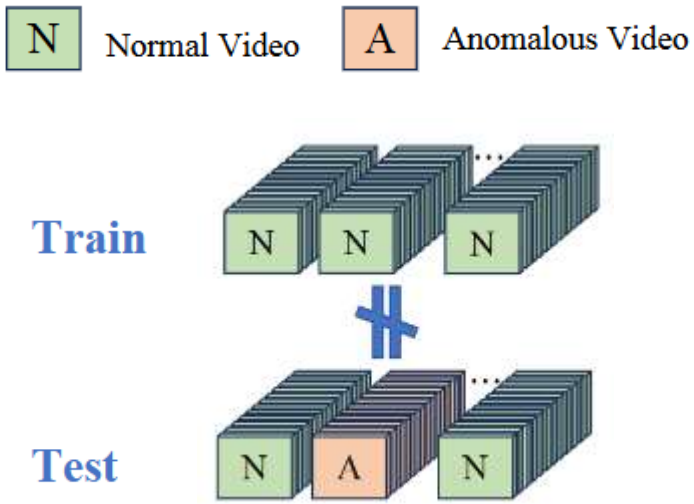
- Background
  - What is Video anomaly detection(VAD)?
    - ANOMALY(이상)은 표준적이거나 정상적이거나 기대되는 것에서 벗어나는 것을 의미
    - 이 중에서도 비디오 이상 탐지(VAD, Video Anomaly Detection)는 중요한 위치를 차지하며, 여기서 이상(anomaly)은 시간적 또는 공간적 차원에서 발생하는 비정상적인 사건을 의미한다.
    - 즉 Video Anomaly Detection은 시간적 또는 공간적 차원에서 발생하는 비정상적인 사건을 탐지해내는 task를 의미

- Background
  - Kind of VAD tasks
    - Train 단계에 어떤 input을 넣느냐에 따라, 크게 5가지 task로 분류가능



CVSP Lab
Computer Vision & Signal Processing

- Background
  - Semi-supervised VAD
    - VAD의 task중, Semi-supervised VAD를 살펴볼 예정
    - Semi-supervised VAD: train때 normal video로만 학습하여 정상데이터의 패턴을 익힌후, test시 normal video 또는 anomalous video를 입력하여 anomaly 여부를 판단하게 하는 task
    - 그중에서도 앞서 제시한 3가지 논문을 살펴볼 예정



Paper review 1

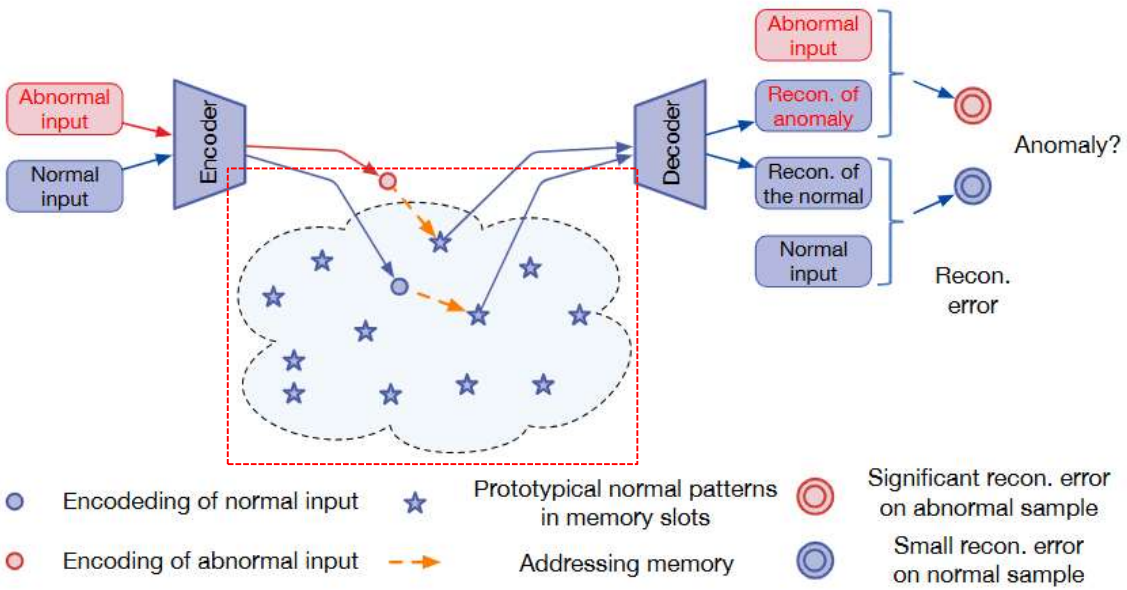
Memorizing Normality to Detect Anomaly: Memory-augmented DeepAutoencoder for Unsupervised Anomaly Detection

ICCV 2019

1) 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.

MemAE

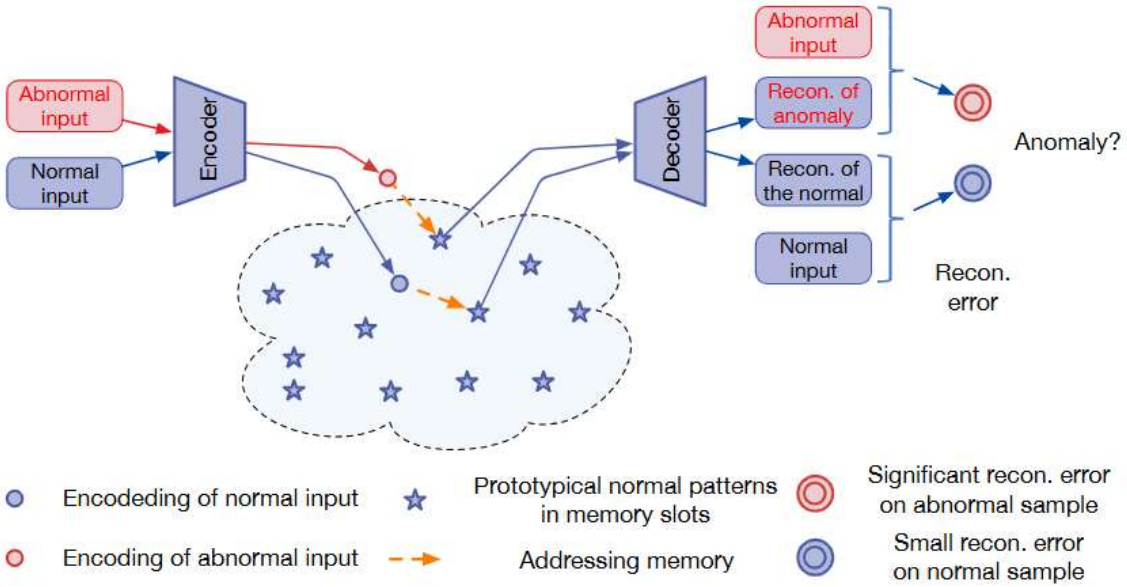
- **Introduction**
  - Memory 기반 복원
    - 기존 Deep autoencoder (AE)는 종종 너무 잘 "일반화"하여 anomaly input도 잘 복원됨
    - Deep autoencoder (AE)의 encoder와 decoder 사이에 memory slot을 추가하여 “너무 좋은” 일반화성능의 문제해결



CVSP Lab Computer Vision & Signal Processing

MemAE

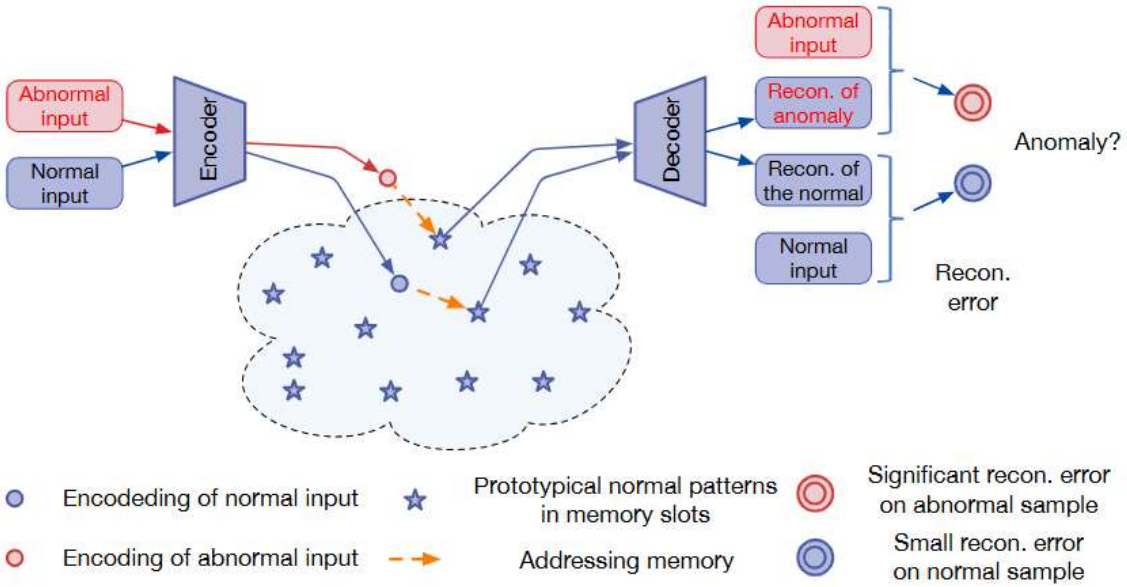
- **Introduction – test시 task의 전체적인 흐름**
  - 그림소개
    - : 인코딩된 normal/abnormal input
    - : memory slot안에 저장된 정상패턴
    - : normal/abnormal 복원 오류값 (input-output)
    - : 인코딩된 input과 가장 유사한 정상패턴을 불러오는 과정



CVSP Lab
Computer Vision & Signal Processing

MemAE

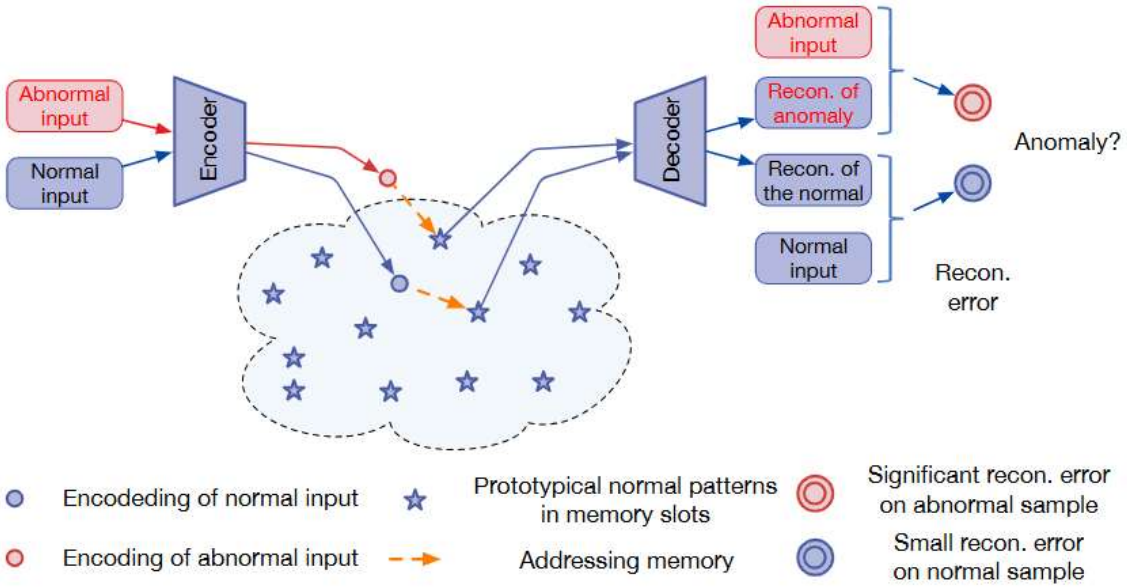
- **Introduction – test시 task의 전체적인 흐름**
  - Train이 끝난 후, Encoding된 input으로 memory slot에 저장된 normal pattern을 address → decoder에 넣어서 input 복원 → 복원된 결과물이 input과 얼마나 유사한지를 판단 → normal or abnormal 판단



CVSP Lab
Computer Vision & Signal Processing

MemAE

- **Introduction – test시 task의 전체적인 흐름**
  - 논문의 논리: “normal input은 normal pattern으로 잘 복원이 되어서 error가 낮을 것이고 abnormal input은 normal pattern으로는 복원이 잘 안될 것이기 때문에 error가 크게 나올 것이다.”

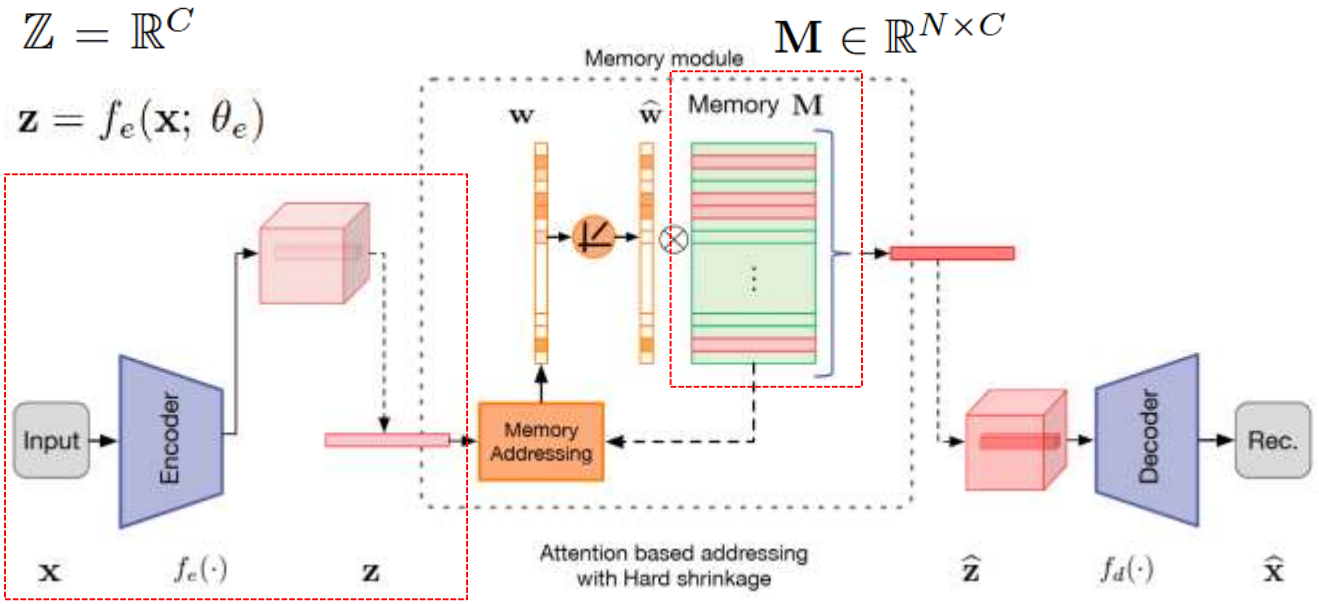


CVSP Lab
Computer Vision & Signal Processing

1) 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.

MemAE

- **Method**
  - X를 인코더에 입력하여 인코딩된 벡터 z를 얻음
  - Memory는 N x C matrix로 구성되어 있고, 각 행 mi, (∀i ∈ [N])이 memory item



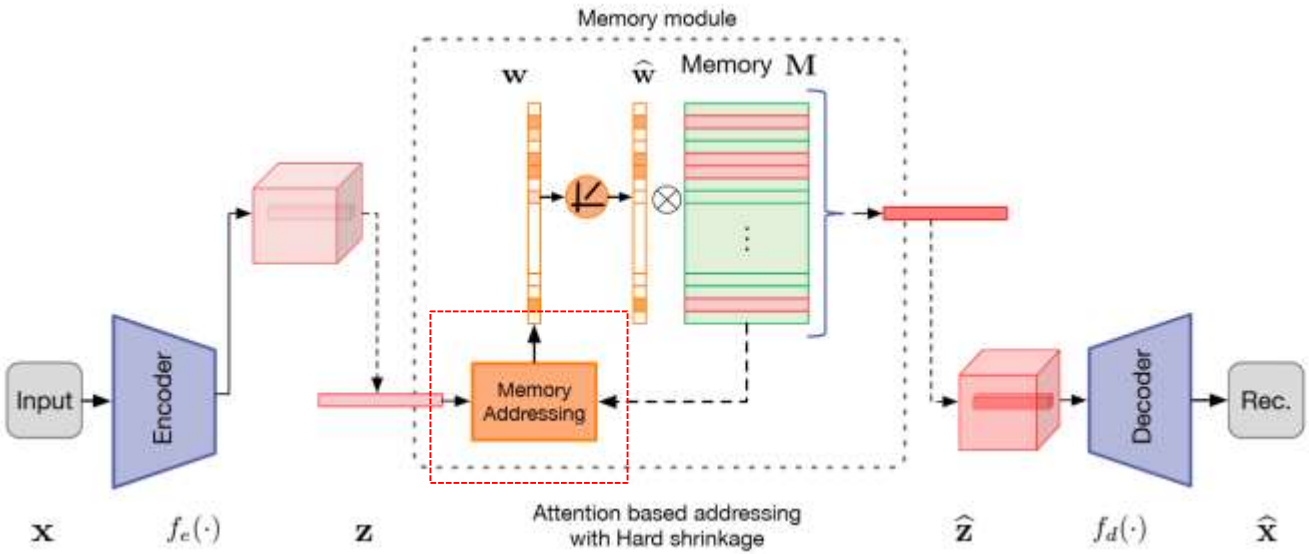
CVSP Lab Computer Vision & Signal Processing

1) 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.

MemAE

- **Method**
  - z와 memory item간 cosine similarity를 계산하여 d를 구하고, 각 memory item의 d에 대한 softmax 계산을 하여 w를 구한다.

$$d(\mathbf{z}, \mathbf{m}_i) = \frac{\mathbf{z}\mathbf{m}_i^{\mathsf{T}}}{\|\mathbf{z}\|\|\mathbf{m}_i\|} \qquad w_i = \frac{\exp(d(\mathbf{z}, \mathbf{m}_i))}{\sum_{j=1}^{N} \exp(d(\mathbf{z}, \mathbf{m}_j))}$$



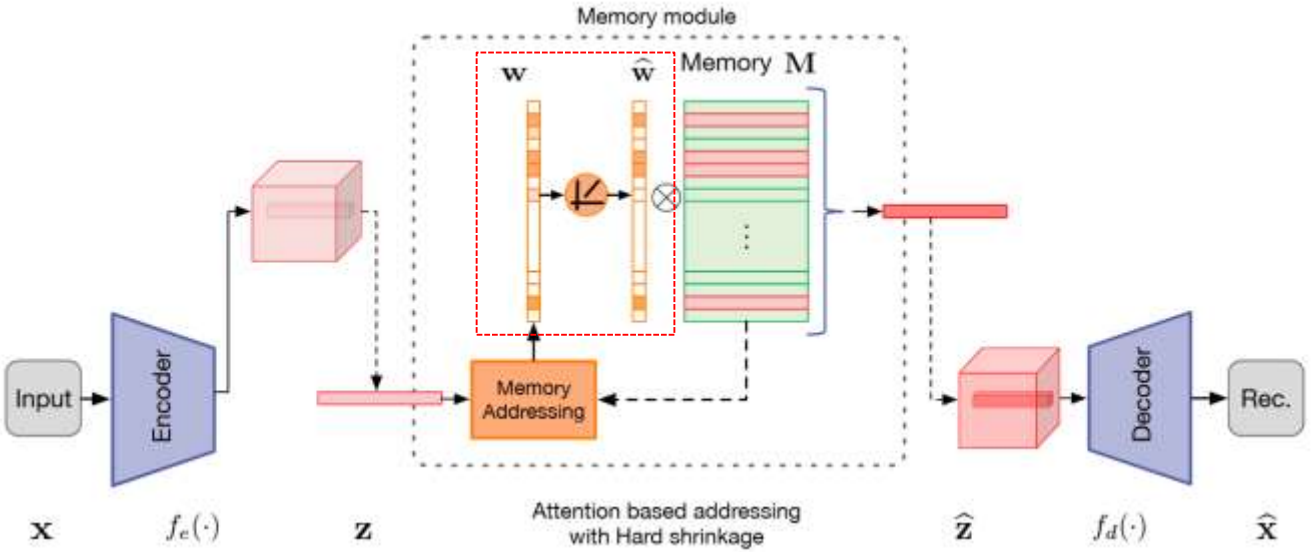
CVSP Lab Computer Vision & Signal Processing

1) 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.

MemAE

- **Method**
  - 구한 w에 대해, 임계값을 넘는 wi만 의미있는 유사도를 가진다 판단하고 나머지 wi는 0으로 만든다

$$\widehat{w}_i = h(w_i; \lambda) = \begin{cases} w_i, & \text{if } w_i > \lambda, \\ 0, & \text{otherwise,} \end{cases}$$



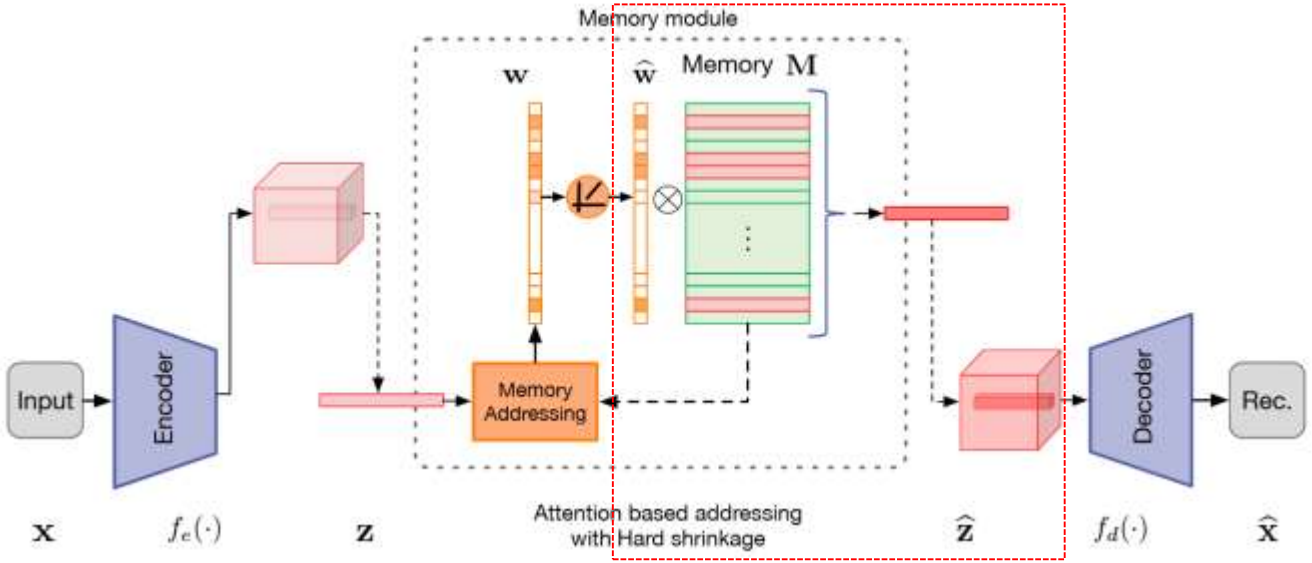
CVSP Lab Computer Vision & Signal Processing

1) 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.

MemAE

- **Method**
  - 남은 w와 memory item간 가중합으로 z를 복원한 $\hat{\mathbf{z}}$ 를 구한다.

$$\hat{\mathbf{z}} = \mathbf{w}\mathbf{M} = \sum_{i=1}^{N} w_i \mathbf{m}_i$$



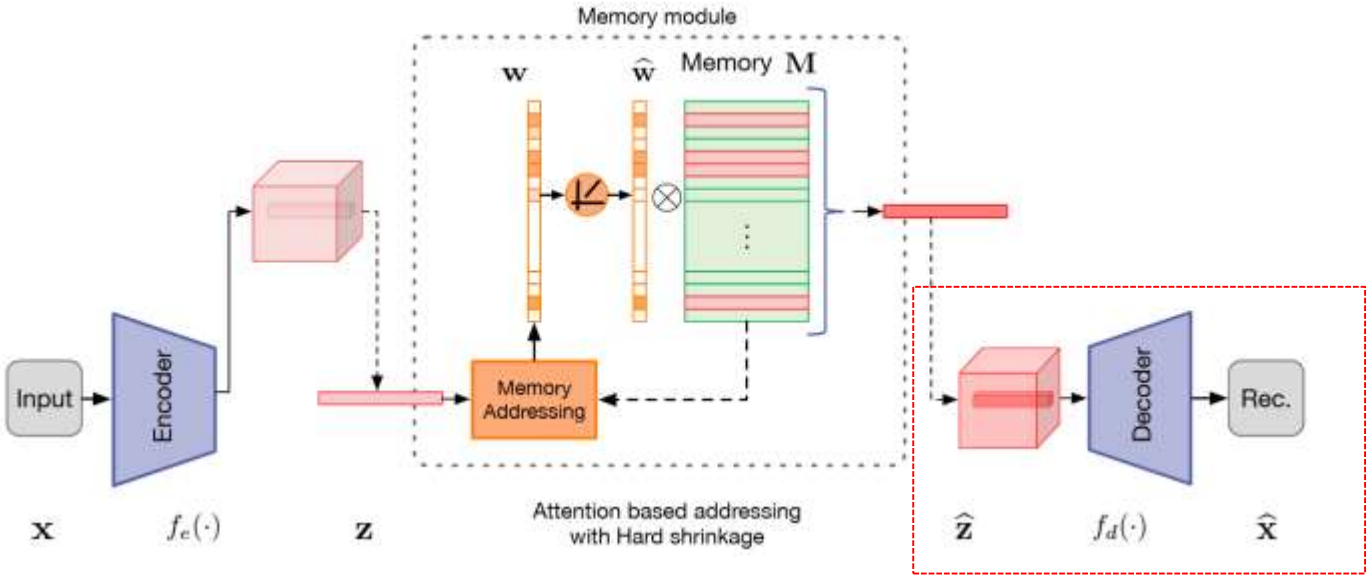
CVSP Lab Computer Vision & Signal Processing

1) 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.

MemAE

- **Method**
  - Decoder를 거쳐서 최종적으로 $\hat{\mathbf{x}}$ 를 복원

$$\hat{\mathbf{x}} = f_d(\hat{\mathbf{z}};\ \theta_d)$$



CVSP Lab Computer Vision & Signal Processing

MemAE

1) 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.

- **Method**
  - Frame level
    - Input frame 하나에 대해, normal/abnormal 판단

1) 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.

MemAE

- **Method**
  - Video level
    - 16개의 grayscale frame을 cube 형태로 한번에 받음 → encoder, decoder 거친 후 16개 각각에 대한 anomaly score 계산 → 모든 프레임의 최종 스코어는 본인을 중심으로 갖는 cube상의 anomaly score
      - ex) 8번째 프레임의 스코어 = 1~16까지의 프레임으로 이루어진 cube를 encode → decode 한 후의 8번째 프레임의 스코어
        - ✓간략화 한 식: Score of f8 : cube (1~16) → encode → decode → cube상의 f8의 anomaly score
        - ✓f9: 2~17 → f9, f10: 3~18 → f10 ….
    - 논문인용(1710p)
      - “the input of the network is a cuboid constructed by stacking 16 neighbor frames in grayscale.”
      - “In testing, the normality of each frame is evaluated by the reconstruction error of the cuboid centering on it.”

CVSP Lab
Computer Vision & Signal Processing

1) 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.

MemAE

- **Method**
  - Loss
    - Reconstruction loss

$$R(\mathbf{x}^t, \hat{\mathbf{x}}^t) = \|\mathbf{x}^t - \hat{\mathbf{x}}^t\|_2^2$$

    - Entropy loss

$$E(\hat{\mathbf{w}}^t) = \sum_{i=1}^{T} -\hat{w}_i \cdot \log(\hat{w}_i).$$

      - 소수의 wi만 1에 가깝고 나머지는 0에 가까우면 loss값이 작아짐
      - 즉 입력을 복원할때 소수의 memory만 선택하게 하여 anomaly 복원시, 큰 error를 유도함(memory 여러개를 할당하면 anomaly input이 들어와도 복원을 잘 해낼 가능성을 사전에 방지)

    - Combined loss

$$L(\theta_e, \theta_d, \mathbf{M}) = \frac{1}{T}\sum_{t=1}^{T}\left(R(\mathbf{x}^t, \hat{\mathbf{x}}^t) + \alpha E(\hat{\mathbf{w}}^t)\right)$$

CVSP Lab
Computer Vision & Signal Processing

MemAE

1) 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.

- **Method**
  - Anomaly score
    - Reconstruction loss와 같은 함수

$$e = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$$

MemAE

- **Experiment**
  - Image

Dataset	MNIST	CIFAR-10
OC-SVM [35]	0.9499	0.5619
KDE	0.8116	0.5756
VAE [18]	0.9643	0.5725
PixCNN [38]	0.6141	0.5450
DSEBM [43]	0.9554	0.5725
AE	0.9619	0.5706
MemAE-nonSpar	0.9725	0.6058
MemAE	**0.9751**	**0.6088**

  - video

	Method\Dataset	UCSD-Ped2	CUHK	SH.Tech
Non-Recon.	MPPCA [15]	0.693	-	-
	MPPCA+SFA [27]	0.613	-	-
	MDT [27]	0.829	-	-
	AMDN [41]	0.908	-	-
	Unmasking [37]	0.822	0.806	-
	MT-FRCN [12]	0.922	-	-
	Frame-Pred [26]	**0.954**	**0.849**	**0.728**
Recon.	AE-Conv2D [11]	0.850	0.800	0.609
	AE-Conv3D [45]	0.912	0.771	-
	TSC [26]	0.910	0.806	0.679
	StackRNN [26]	0.922	0.817	0.680
	AE	0.917	0.810	0.697
	MemAE-nonSpar	0.929	0.821	0.688
	MemAE	**0.941**	**0.833**	**0.712**

Paper review 2

Learning Memory-guided Normality for Anomaly Detection

CVPR 2020

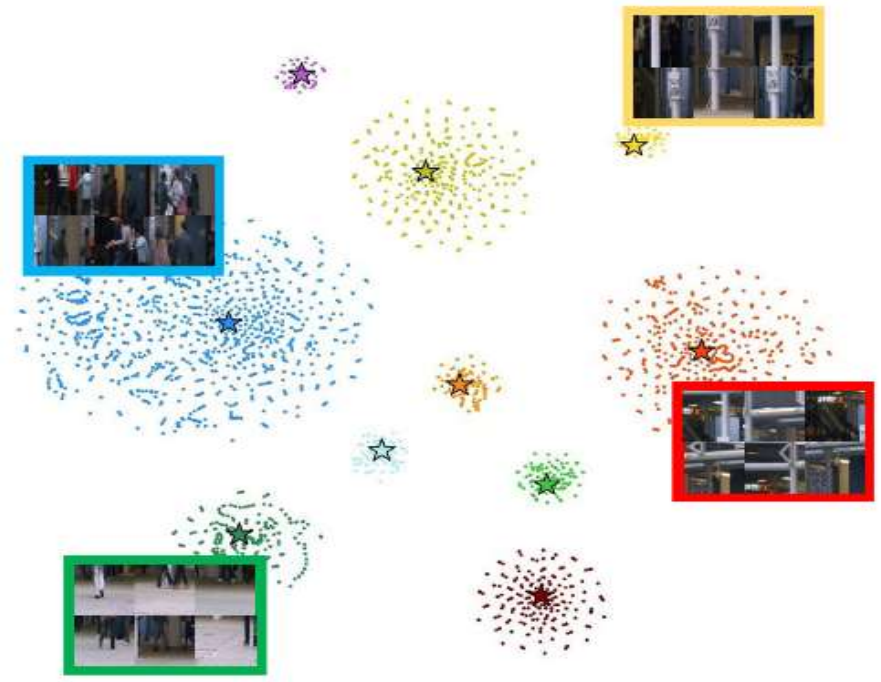
1) Park, Hyunjong, Jongyoun Noh, and Bumsub Ham. "Learning memory-guided normality for anomaly detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

- **Introduction**

  - 기존에 있던 MemAE는 정상패턴의 다양성을 고려하지 못함 그래서 memory 용량이 너무 큼

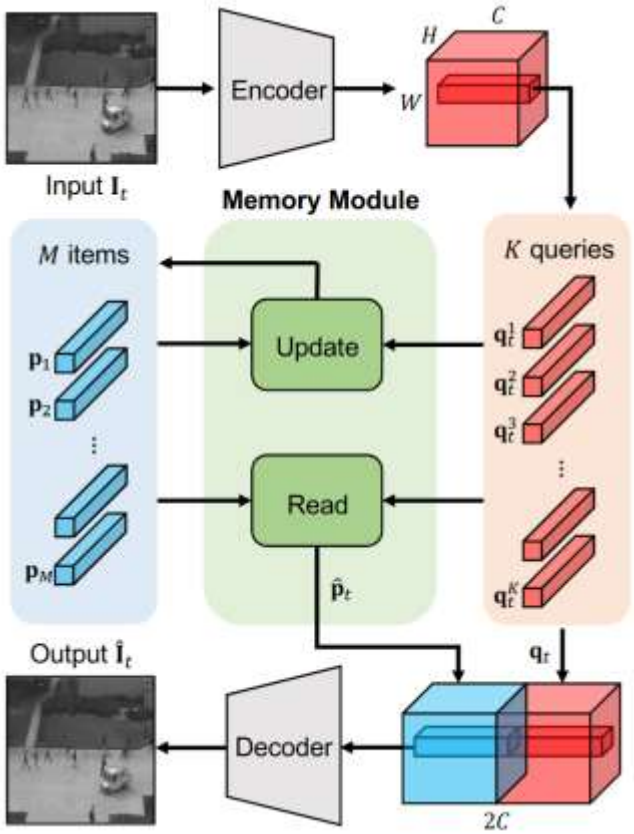
  - 이를 해결하기 위해 compact loss, separate loss, test시에도 **실시간 update**를 도입하여 정상패턴의 다양성 보장 및 용량 문제 해결

- **Introduction**
  - 별: 정상패턴의 다양성을 고려한 memory item
  - 점: 입력으로 들어온 데이터



MNAD

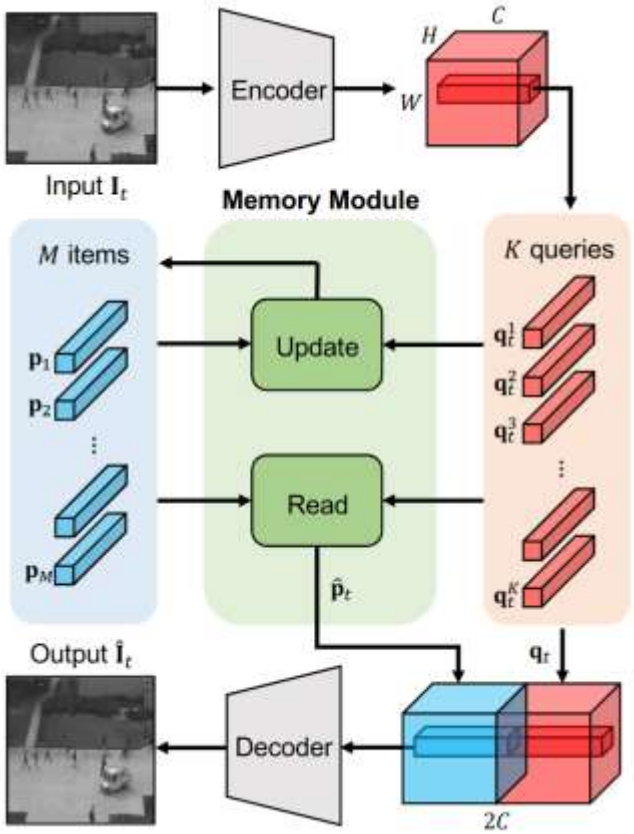
- **Method**
  - $\mathbf{q}_t^k$: input query
    - size: 1x1xc, (k = 1, . . . K), K = HxW
  - $\mathbf{p}_m$: memory item
    - size: 1x1xc, (m = 1, . . . , M)



CVSP Lab
Computer Vision & Signal Processing

MNAD

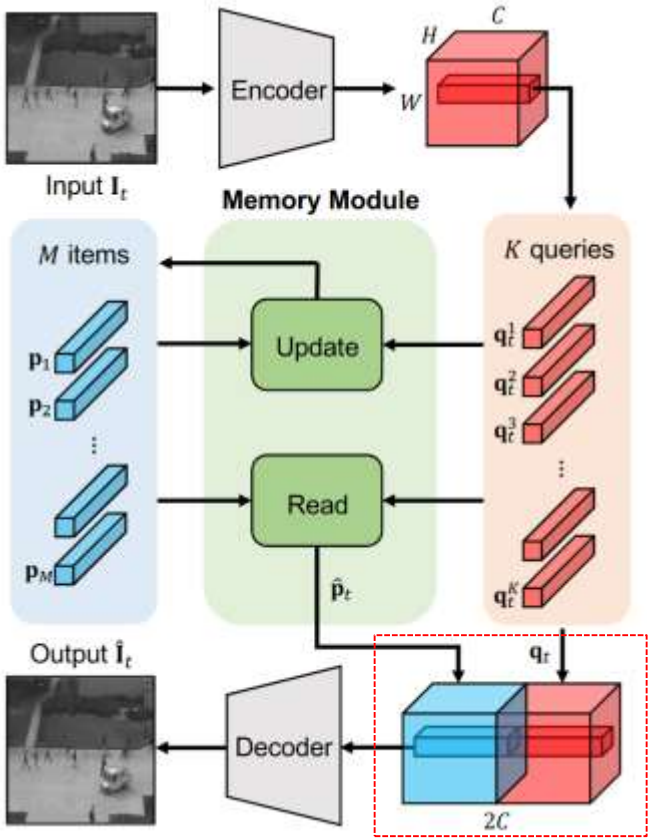
- **Method**
  - Read: 입력 q를 memory item인 pm으로 재구성하는 과정
  - Update: memory item p를 입력인 q를 반영하여 update 하는 과정



CVSP Lab
Computer Vision & Signal Processing

MNAD

- **Method**
  - Concatenation: CNN의 anomaly input의 일반화성능 제한
    - Normal input: p,q 차이안남 → 복원할때 수월
    - Abnormal Input: p,q 차이 많이남 → 복원할때 제대로 복원x

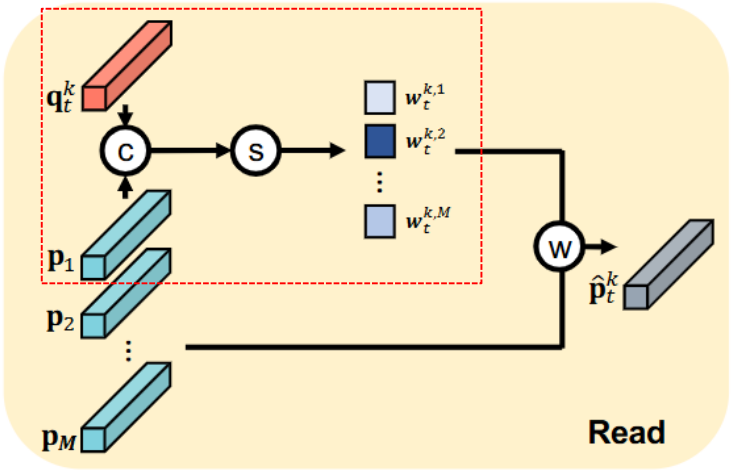


CVSP Lab
Computer Vision & Signal Processing

MNAD

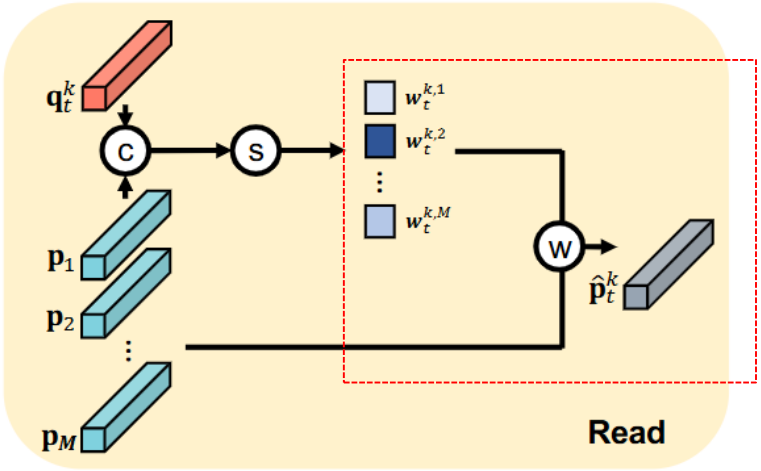
- **Method**
  - read
    - $\mathbf{q}_t^k$ 와 모든 memory item 간 cosine similarity를 구한 뒤, softmax 계산으로 m번째 memory item에 대한 weight를 모든 m에 대해 구한다.

$$w_t^{k,m} = \frac{\exp((\mathbf{p}_m)^T \mathbf{q}_t^k)}{\sum_{m'=1}^{M} \exp((\mathbf{p}_{m'})^T \mathbf{q}_t^k)}$$



- **Method**
  - read
    - 구한 w들과 pm간에 weighted sum을 통해 $\mathbf{q}_t^k$를 pm에 대한 가중합으로 최종 표현한 것을 볼 수 있다.

$$\hat{\mathbf{p}}_t^k = \sum_{m'=1}^{M} w_t^{k,m'} \mathbf{p}_{m'}$$

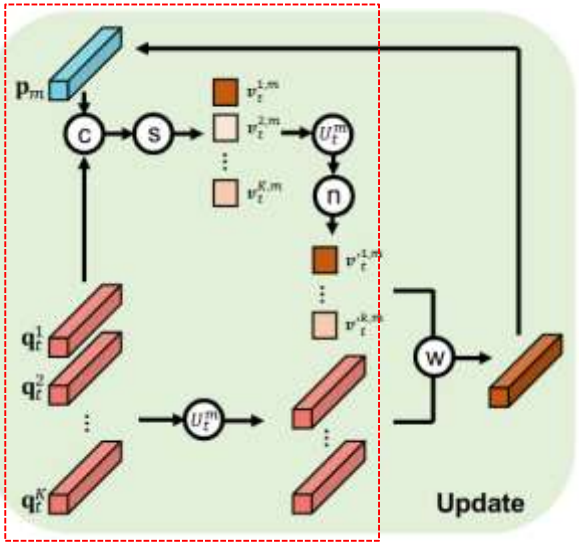


CVSP Lab
Computer Vision & Signal Processing

MNAD

- **Method**
  - Update
    - Pm에 대해, 모든 qt들과의 cosine similarity를 구한 후, softmax로 weight를 계산
    - V에 대해, 큰 비중을 차지하는 v를 update에 더 많은 영향을 끼치기위해 max인 v로 전체 v를 나눠준다(renormalize).
    - 여기서 U_t^m은 m번째 메모리항목과 관련있는 쿼리들의 인덱스 집합이라 한다.

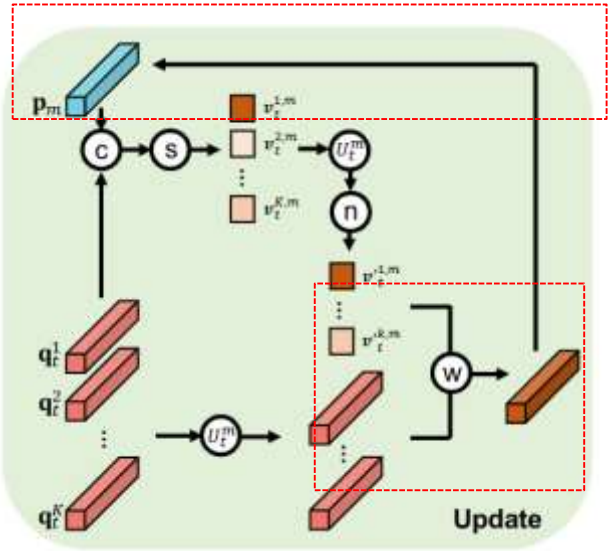
$$v_t^{k,m} = \frac{\exp((\mathbf{p}_m)^T \mathbf{q}_t^k)}{\sum_{k'=1}^{K} \exp((\mathbf{p}_m)^T \mathbf{q}_t^{k'})} \qquad v'^{k,m}_t = \frac{v_t^{k,m}}{\max_{k' \in U_t^m} v_t^{k',m}}$$



MNAD

- **Method**
  - Update
    - 가중치를 반영하여 입력쿼리 q들의 가중합으로 pm을 update한다.
    - 연산에 U_t^m이 들어가는 이유는 pm을 업데이트할때, pm과 관련있는 즉 거리가 가까운 쿼리들의 정보를 반영하여 업데이트 하는것이다.

$$\mathbf{p}^m \leftarrow f(\mathbf{p}^m + \sum_{k \in U_t^m} v'^{k,m}_t \mathbf{q}^k_t)$$



MNAD

- **Method**
  - Update
    - Test 시에도 update 계속 진행
      - regular score Et를 이용해 Et의 값이 threshold γ를 넘을경우, input을 anomaly로 간주하고, update에 반영하지 않음
      - 이 과정을 통해 test 환경에 맞춰 memory item을 update → 소수의 memory item으로도 normal/abnormal 판단가능

$$W_{ij}(\hat{\mathbf{I}}_t, \mathbf{I}_t) = \frac{1 - \exp(-||\hat{\mathbf{I}}_t^{ij} - \mathbf{I}_t^{ij}||_2)}{\sum_{i,j} 1 - \exp(-||\hat{\mathbf{I}}_t^{ij} - \mathbf{I}_t^{ij}||_2)}$$

$$\mathcal{E}_t = \sum_{i,j} W_{ij}(\hat{\mathbf{I}}_t, \mathbf{I}_t) \|\hat{\mathbf{I}}_t^{ij} - \mathbf{I}_t^{ij}\|_2$$

CVSP Lab
Computer Vision & Signal Processing

MNAD

- **Method**
  - Loss

$$\mathcal{L}_{\text{rec}} = \sum_{t}^{T} \|\hat{\mathbf{I}}_t - \mathbf{I}_t\|_2$$

- Reconstruction loss: 원본이미지와 복원한 이미지의 차이를 계산하는 loss

$$\mathcal{L}_{\text{compact}} = \sum_{t}^{T} \sum_{k}^{K} \|\mathbf{q}_t^k - \mathbf{p}_p\|_2$$

- Feature compactness loss: 쿼리와 memory item간 거리를 좁히는 loss

$$\mathcal{L}_{\text{separate}} = \sum_{t}^{T} \sum_{k}^{K} [\|\mathbf{q}_t^k - \mathbf{p}_p\|_2 - \|\mathbf{q}_t^k - \mathbf{p}_n\|_2 + \alpha]_+$$

- Feature separateness loss: 각 쿼리에 대해, 가장 가까운 memory item과는 가까워지고, 두번째로 가까운 memory item과는 멀어지게 하는 loss

CVSP Lab Computer Vision & Signal Processing

MNAD

- **Method**
  - Loss
    - Pp: 어떤 쿼리와 feature space상에서 가장 거리가 가까운 memory item
    - Pn: 어떤 쿼리와 feature space상에서 두번째로 거리가 가까운 memory item

$$\mathcal{L}_{\text{compact}} = \sum_{t}^{T} \sum_{k}^{K} \|\mathbf{q}_t^k - \mathbf{p}_p\|_2$$

- Feature compactness loss: 쿼리와 가장 가까운 memory item 간 거리를 좁히는 함수

$$\mathcal{L}_{\text{separate}} = \sum_{t}^{T} \sum_{k}^{K} [\|\mathbf{q}_t^k - \mathbf{p}_p\|_2 - \|\mathbf{q}_t^k - \mathbf{p}_n\|_2 + \alpha]_+$$

- Feature separateness loss: 각 쿼리에 대해, 가장 가까운 memory item과는 가까워지고, 두번째로 가까운 memory item과는 멀어지게 하는 loss

MNAD

- **Method**
  - Loss

– Compact loss, separate loss, test time update를 통해, 불필요한 memory item을 줄이고, 필요한 memory item만 남김으로써, 메모리용량을 획기적으로 줄임

– “Note that MemAE exploits 3D convolutions with **2,000** memory items of size 256. On the contrary, our model uses 2D convolutions and it requires **10** items of size 512.”(논문내용)

MNAD

- **Method**
  - Anomaly score
    - Test 시 D, P를 구한후, g로 min-max normalization후 최종 S를 계산

$$D(\mathbf{q}_t, \mathbf{p}) = \frac{1}{K}\sum_{k}^{K} \|\mathbf{q}_t^k - \mathbf{p}_p\|_2 \qquad P(\hat{\mathbf{I}}_t, \mathbf{I}_t) = 10\log_{10}\frac{\max(\hat{\mathbf{I}}_t)}{\|\hat{\mathbf{I}}_t - \mathbf{I}_t\|_2^2/N.}$$

$$g(D(\mathbf{q}_t, \mathbf{p})) = \frac{D(\mathbf{q}_t, \mathbf{p}) - \min_t(D(\mathbf{q}_t, \mathbf{p}))}{\max_t(D(\mathbf{q}_t, \mathbf{p})) - \min_t(D(\mathbf{q}_t, \mathbf{p}))}$$

$$\mathcal{S}_t = \lambda(1 - g(P(\hat{\mathbf{I}}_t, \mathbf{I}_t))) + (1-\lambda)g(D(\mathbf{q}_t, \mathbf{p}))$$

CVSP Lab
Computer Vision & Signal Processing

MNAD

- **Experiment**
  - 성능자체는 MemAE를 뛰어넘지는 못했지만, memory item을 획기적으로 줄이는 방식을 선보임으로써, memory 기반 VAD 분야에 지대한 영향을 끼친 논문이라 생각함

	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
Recon.	TSC [26]	91.0	80.6	67.9
Recon.	StackRNN [26]	92.2	81.7	68.0
Recon.	AbnormalGAN [33]	93.5	-	-
Recon.	MemAE w/o Mem. [8]	91.7	81.0	69.7
Recon.	MemAE w/ Mem. [8]	94.1	83.3	71.2
Recon.	Ours-R w/o Mem.	86.4	80.6	65.8
Recon.	Ours-R w/ Mem.	90.2	82.8	69.8
Pred.	Frame-Pred [22]	95.4	85.1	**72.8**
Pred.	Ours-P w/o Mem.	94.3	84.5	66.8
Pred.	Ours-P w/ Mem.	**97.0**	**88.5**	70.5

CVSP Lab
Computer Vision & Signal Processing

Paper review 3

Video Anomaly Detection by Solving Decoupled Spatio-Temporal Jigsaw Puzzles

ECCV 2022

1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Introduction**
  - 기존의 방식과 전혀 다른 문제해결 방식
  - 공간 및 시간 차원으로 분리된 퍼즐 구성
  - 멀티레이블 학습(multi-label learning) 방식 적용

1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Introduction**
  - 기존의 방식과 전혀 다른 문제해결 방식

    - 사전 학습된 네트워크를 전혀 사용하지 않으면서도 SOTA를 달성

    - 논문의 논리: “Train때 순서를 맞춰 나가는 과정에서 normal 프레임의 공간적 시간적 정상패턴을 학습하게 될 것이고 그것을 바탕으로 test때 normal/abnormal을 구분할 수 있을 것이다.”

STJP

1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Introduction**
  - 공간 및 시간 차원으로 분리된 퍼즐 구성

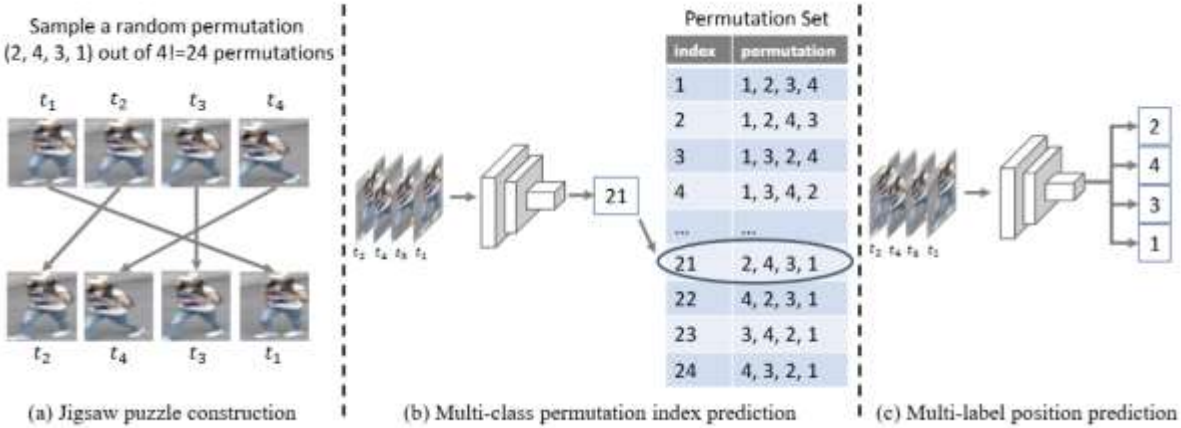
- **공간적 직소 퍼즐(spatial jigsaw puzzles)** : 개별 프레임 내에서 패치를 무작위로 섞음 → 정상적인 외형(appearance) 패턴 학습

- **시간적 직소 퍼즐(temporal jigsaw puzzles)** : 연속 프레임의 순서를 무작위로 변경 → 정상적인 동작(motion) 패턴 학습

1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Introduction**
  - 멀티레이블 학습(multi-label learning) 방식 적용

    - 기존 연구에서는 각 순열을 개별 클래스로 간주하여 다중 클래스 분류 문제로 정의했으나, 이는 메모리 소비가 크다는 단점
    - 본 연구에서는 직소 퍼즐 해결을 **멀티레이블 학습 문제로 변환**하여 확장 가능성을 높이고 메모리 소비를 줄임



1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

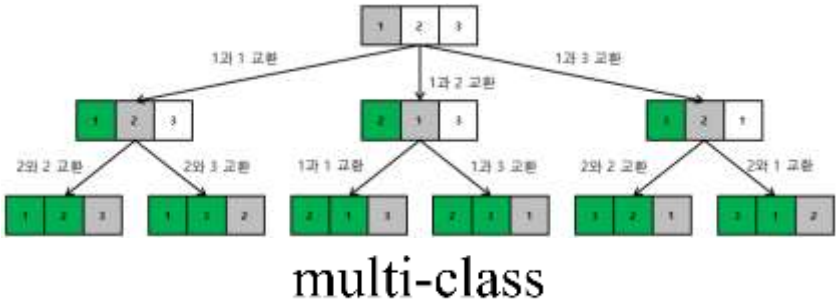
- **Introduction**
  - 멀티레이블 학습(multi-label learning) 방식 적용

– 멀티클래스 계산: n개의 퍼즐 → n!가지 경우를 다 따져야함 multi-class

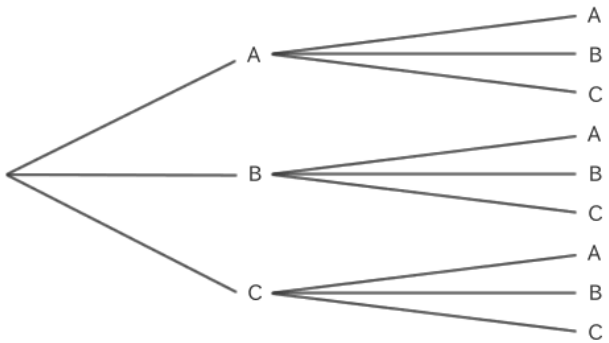
⁘ 계산량: $n! = n \times (n-1) \times \cdots \times 2 \times 1$ → 모든 순열의 경우의 수 다 따져야 함

– 멀티레이블 계산: n개의 퍼즐 → n^2가지 경우만 따지면됨

⁘ 계산량: n^2 → n이 클수록 계산량이 멀티클래스보다 현저히 줄어듦



multi-class

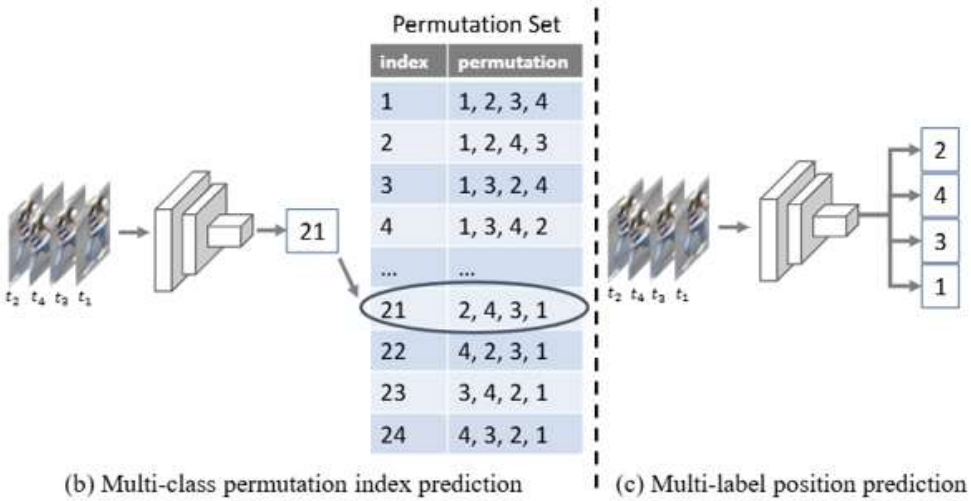


multi-label

1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Introduction**
  - 멀티레이블 학습(multi-label learning) 방식 적용

- 기존 멀티 클래스: n개의 패치 또는 프레임에 대해, n!개의 경우의 수를 따져야함 $O(n!)$
  → 연산이 너무 많아짐
- 멀티 레이블: n개의 패치 또는 프레임에 대해, n의 제곱개의 경우의 수로 줄어듦 $O(n^2)$
  → 연산 줄어듦
- 이 전략은 문제의 복잡도를 $O(n!)$에서 $O(n^2)$로 줄이며, 메모리 소비가 미미하다.



(b) Multi-class permutation index prediction (c) Multi-label position prediction

STJP

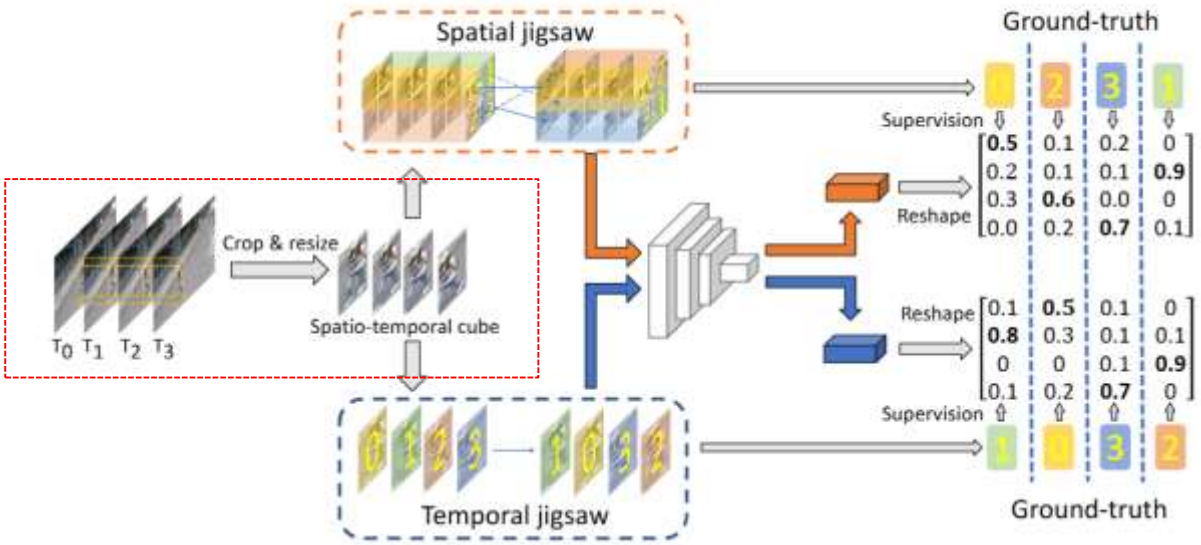
1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Method**
  - object-centric cube extraction
  - puzzle construction
  - puzzle solving

STJP

1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Method**
  - object-centric cube extraction
    - Yolov3를 이용하여 먼저 각 프레임에서 object만 따로 추출한다.

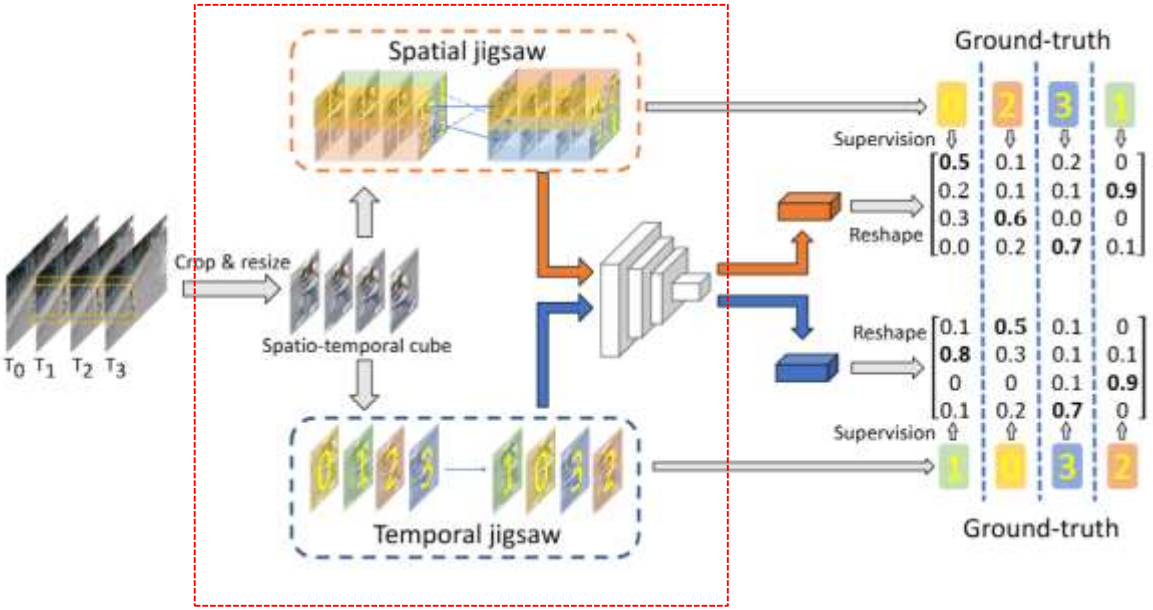


CVSP Lab
Computer Vision & Signal Processing

STJP

1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Method**
  - puzzle construction
    - Spatial: 각 object 프레임에 대해 nxn의 패치로 쪼개어서 순서를 무작위로 섞는다.
    - Temporal: l개의 object 프레임을 하나의 묶음으로 묶어서 그들간의 프레임 순서를 무작위로 섞는다.

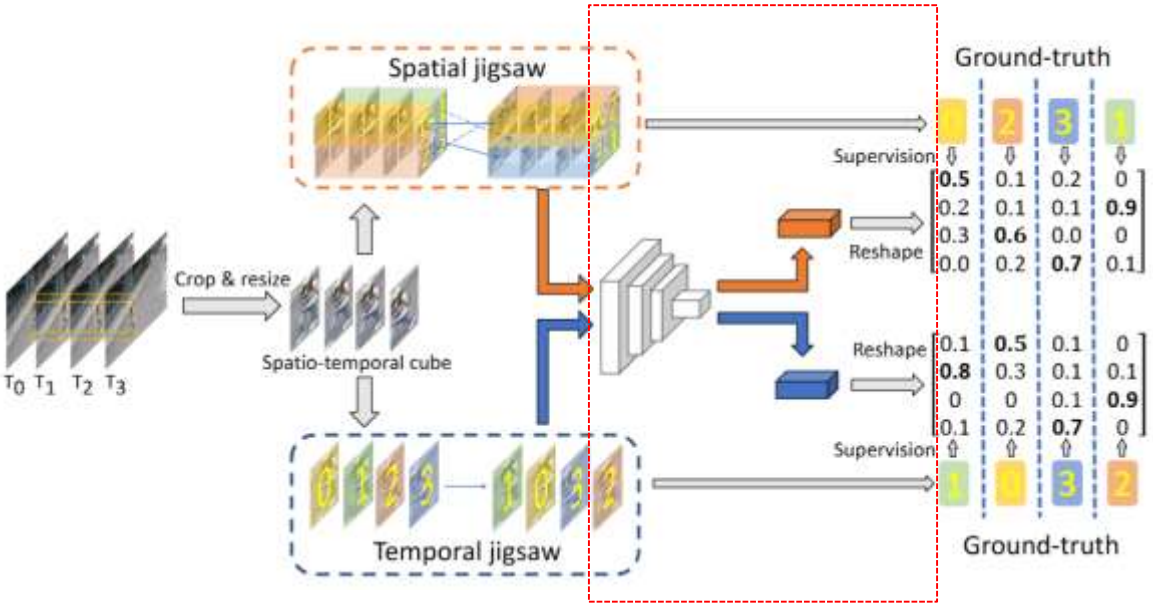


CVSP Lab
Computer Vision & Signal Processing

STJP

1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

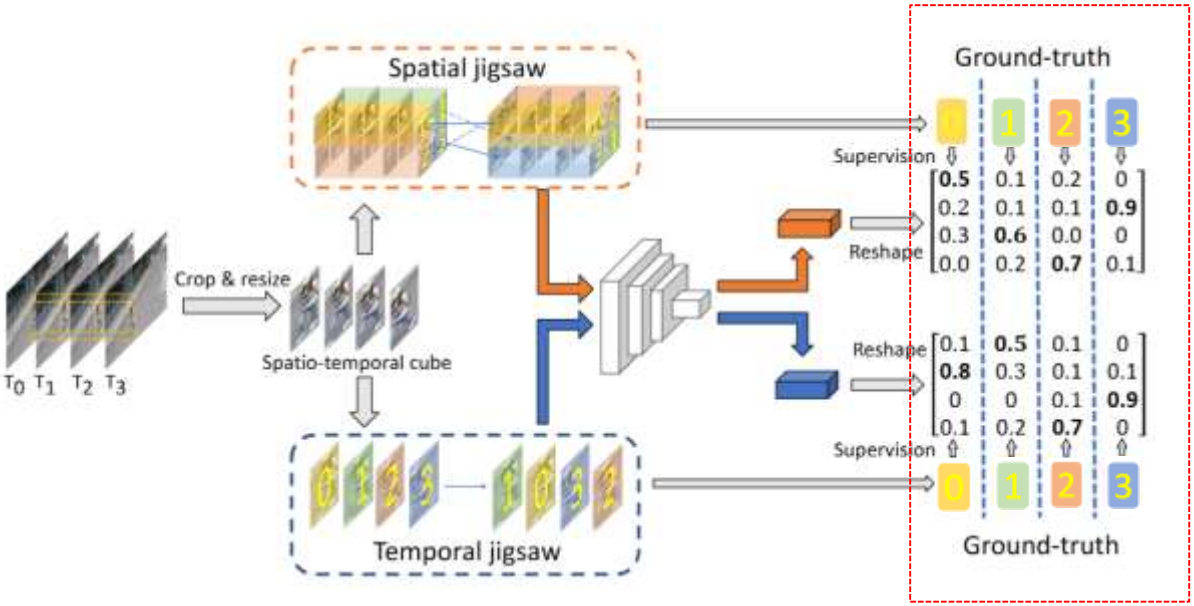
- **Method**
  - puzzle solving
    - 공유된 3D 컨볼루션 백본으로 퍼즐을 보낸다.
    - 백본에서는 공간적 퍼즐과 시간적 퍼즐을 독립적으로 순서를 예측함



CVSP Lab
Computer Vision & Signal Processing

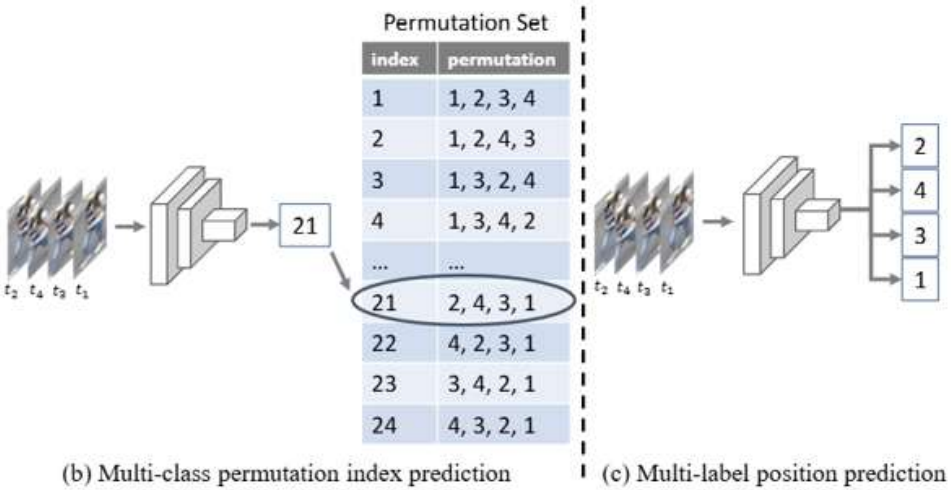
1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Method**
  - puzzle solving
    - 멀티 레이블 분류 문제를 풀어서 ground truth과 예측결과를 비교한다.



1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Method**
  - puzzle solving
    - 기존 멀티클래스 분류의 경우, 모든 n!에 대해 클래스로 분류한 반면, 멀티레이블의 경우, n개의 퍼즐이 n개의 순서를 개별적으로 예측하게 하면 되므로 n^2가지의 경우의 수를 계산하면 된다. 문제의 복잡도 O(n!) → O(n^2)



1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Method**
  - Loss
    - Spatial 과 temporal에 대해 각각 cross entropy loss를 계산한다.

$$L_p = \begin{cases} \frac{1}{l}\sum_{i=1}^{l} CE(t_i, \hat{t}_i), & p \in Q_t \\ \frac{1}{n^2}\sum_{j=1}^{n^2} CE(s_j, \hat{s}_j), & p \in Q_s \end{cases}$$

1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

• Method

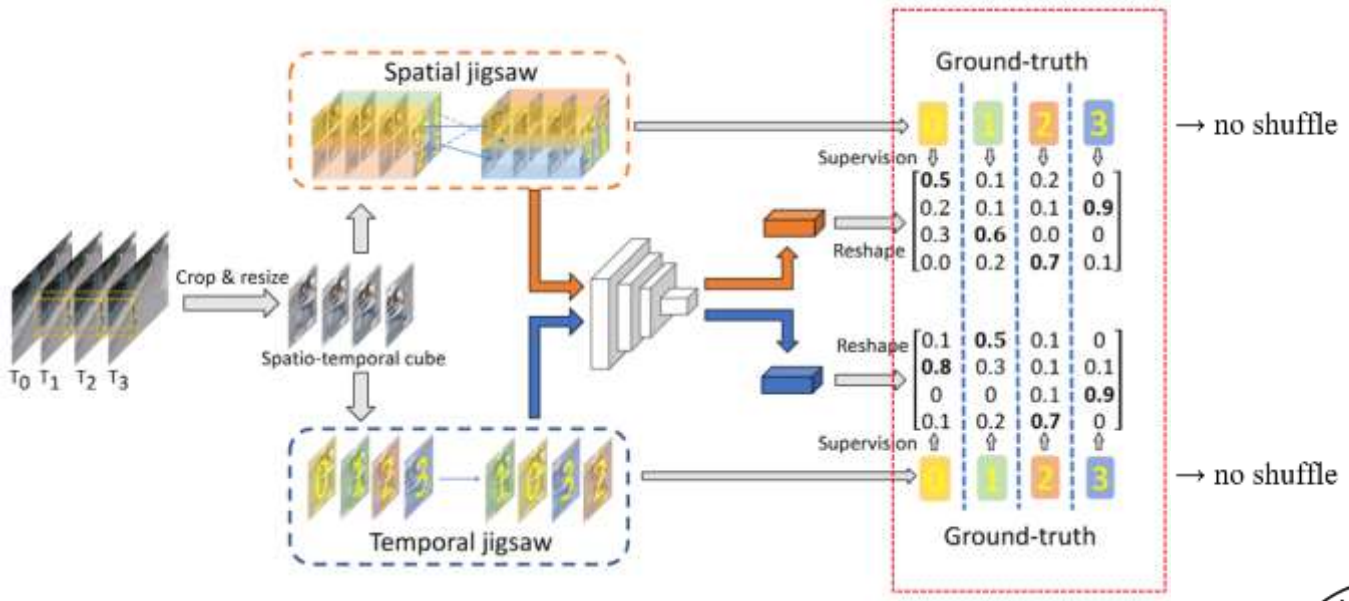
▪ Inference

– Spatial과 temporal에 대해 각각 예측점수 matrix를 만든다. (Ms, Mt)

– 단, inference 시에는 섞기를 하지않고 그대로 puzzle solving 실행

– 이 matrix의 대각선에 각 프레임에 대해 예측을 잘했는지가 나옴

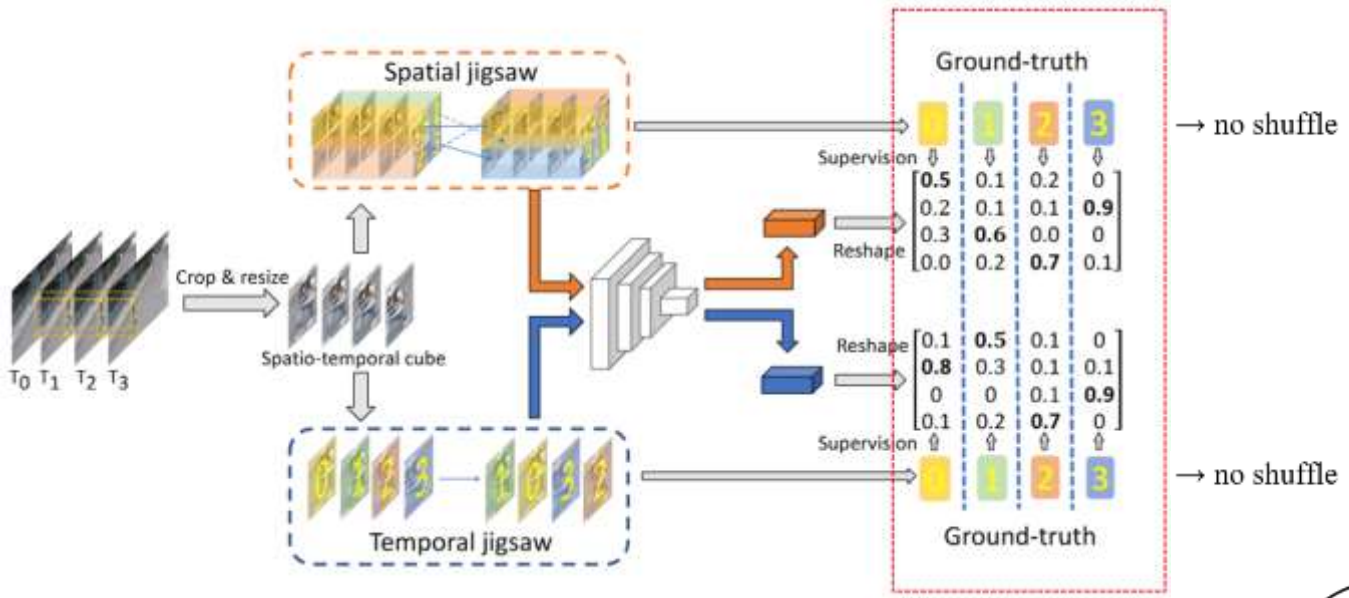
- 행(row): 예측된 패치의 위치
- 열(column): 실제 패치의 원래 위치(ground truth)
- 대각선(diagonal)이 클수록 정상(normal) 데이터일 가능성이 높음



1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Method**
  - Inference
    - 대각선 성분 중 가장 작은값을 regularity score로 선정
    - Score가 높을수록 normal에 가깝다는 의미

$$\begin{cases} r_s = \min(\mathrm{diag}(M_s)) \\ r_t = \min(\mathrm{diag}(M_t)) \end{cases}$$



CVSP Lab
Computer Vision & Signal Processing

STJP

1) Wang, Guodong, et al. "Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

- **Experiment**
  - **Denotes that micro-AUROC is reported.**

Type	Method	Ped2	Avenue	STC
reconstruction	Conv-AE [18]	90.0	70.2	-
	StackRNN [35]	92.2	81.7	68.0
	Mem-AE [16]	94.1	83.3	71.2
	AM-Corr [40]	96.2	86.9	-
	MNAD-Recon. [43]	90.2	82.8	69.8
	ClusterAE [6]	96.5	86.0	73.3
	VEC [59]	97.3	90.2	74.8
	LNRA (Patch based) [4]	94.8	84.9	72.5
	LNRA (Skip frame based) [4]	96.5	84.7	76.0
	I3D-Recons.	-	69.3 -	69.4
prediction	Frame-Pred. [29]	95.4	85.1	72.8
	BMAN [27]	96.6	90.0	76.2
	Multipath-Pred. [52]	96.3	88.3	76.6
	MNAD-Pred. [43]	97.0	88.5	70.5
	CT-D2GAN [56]	97.2	85.9	77.7
	Bi-Prediction [7]	96.6	87.8	-
hybrid	ST-CAE [61]	91.2	80.9	-
	MPED-RNN [38]	-	-	73.4
	AnoPCN [58]	96.8	86.2	73.6
	IntegradAE [50]	96.3	85.1	73.0
	HF^2-VAD [30]	**99.3**	91.1	76.2
others	SCL [32]	-	80.9	-
	DeepOC [55]	96.9	86.6	-
	CAE-SVM* [21]	94.3	87.4	78.7
	Scene-Aware [49]	-	89.6	74.7
SSL	CAC [53]	-	87.0	79.3
	SS-MTL* [13]	97.5	91.5	82.4
	Ours	99.0	**92.2**	**84.3**

Conclusion

- Memory: normal video의 feature를 저장해놓은 상태에서 입력이 들어오면 그 입력과 유사한 feature들을 불러와 복원하는 방식(좀더 전형적인 방식)

- Jigsaw puzzle: 들어온 데이터에 대해 퍼즐을 복원하는 방식(획기적이고 새로운 방식)

CVSP Lab
Computer Vision & Signal Processing

Conclusion

- Memory와 jigsaw puzzle은 전혀 다른 접근방식이지만 같은 VAD task에 쓰인다는 것이 인상깊었고, 이 둘을 합쳐서 좋은 성능을 내게 할 수 있지 않을까 하는 생각이 들어 발표를 준비함

- 앞으로 더 알아볼 예정