Video Anomaly Detection with Synthetic Dataset

#VideoAnomalyDetection(VAD) #SyntheticDataset

Presented By 김태훈 학부연구생

2025/02/14

CVSP Lab
Computer Vision & Signal Processing

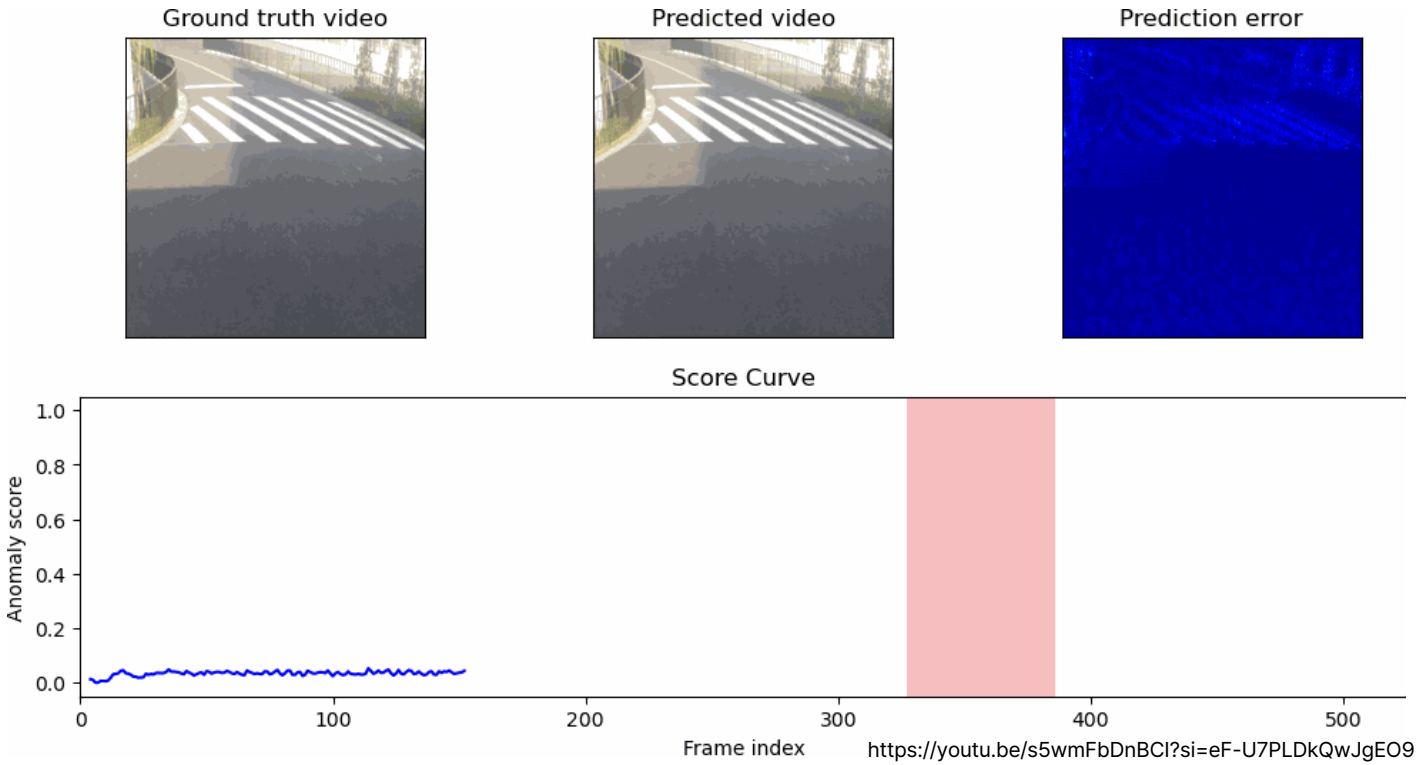
2025 Winter Seminar

Index

1. Video Anomaly Detection Task
2. Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping a.k.a. PFMF
3. Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors a.k.a. AED-MAE

1. Video Anomaly Detection Task

- **Video Anomaly Detection(VAD) Definition**
  - ↳정상적인 패턴에서 벗어나는 화재,낙상과 같은 비정상적인 것들을 탐지함.
  - ↳Input=비디오, Output=이상(Abnormal)여부



1. Video Anomaly Detection Task

- **Characteristics of Abnormal Dataset**
  - ↳매우 적은 빈도로 발생하는 **Unbalanced Dataset**
    - ✓ Normal의 빈도 >> Abnormal의 빈도
  - ↳배경과 맥락에 따라 **Abnormal 여부가 달라질 수 있음.** → e.g. 도로 위를 달리는 차 vs. 도보 위를 달리는 차
    - ✓ Abnormal Labeling의 어려움
  - ∴ 주로 사용되는 데이터셋의 구성
  - ↳Abnormal Annotation: bounding box or pixel level mask
  - ↳Train : normal ONLY (One-Class)
  - ↳Test : normal + abnormal

Dataset Name	Year	Normal Frames↑	Abnormal Frames↑	**Normal/Abnormal↓**	**Synthetic**
CUHK Avenue	2013	26832	3820	**7.02**	**X**
ShanghaiTech	2017	300308	17090	**17.57**	**X**
NWPU Campus	2023	1400807	65266	**21.47**	**X**
UBnormal	2022	147887	89015	**1.66**	**O**

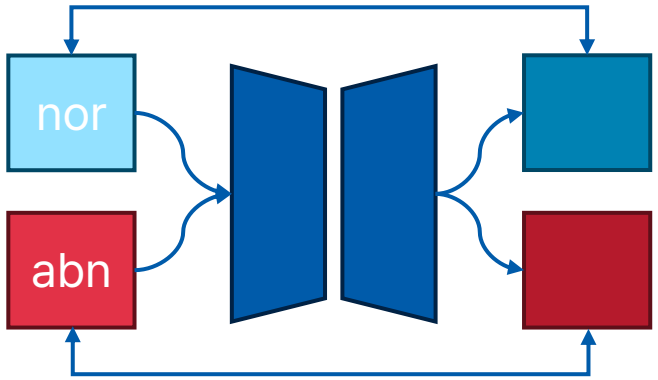


http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

1. Video Anomaly Detection Task

- **Data Driven Solution for VAD**

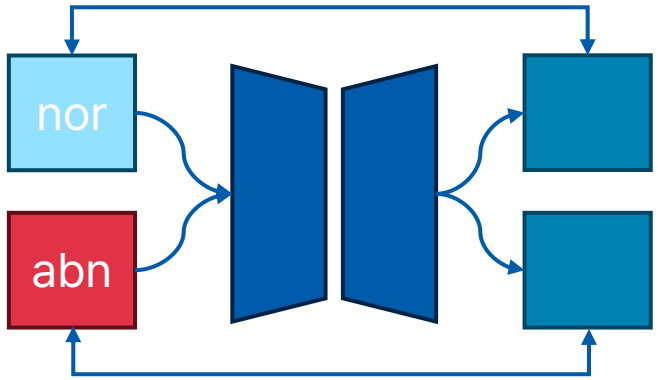
☺ Short Distance



☺ Short Distance

Well Generalized

☺ Short Distance



☹ Long Distance → Anomaly Potential ↑

Overfitting to Normal

↳ "**<u>Normal만 학습하자. 이때, 학습된 패턴에 크게 벗어나는 것을 Abnormal 이라하자.</u>**"

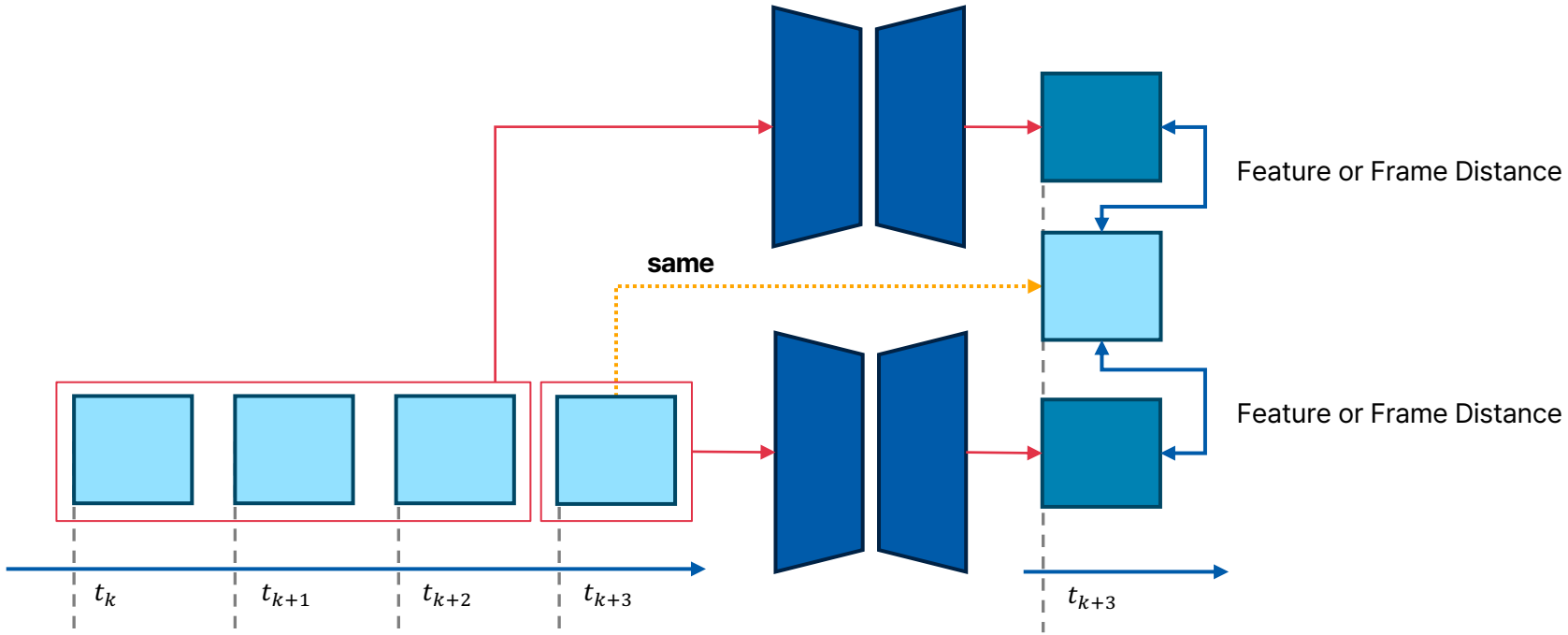
= Normal에 Overfitting 하자

✓ Normal Pattern들을 메모리로 저장 → 메모리들을 짜집기하여 Reconstruction

1. Video Anomaly Detection Task

- **Data Driven Solution for VAD**

  Frame Reconstruction Base, Frame Prediction Base(=reconstruct with past info.)
  두 가지가 많이 쓰임.



↳**BUT** Genaralization capacity of Deep Neural Network(DNN)

✓ Normal만 학습된 DNN이더라도, Abnormal의 Frame을 잘 복원한다.

Liu, Zuhao, et al.

Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping

CVPR 2023

2. PFMF

Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping

- **UBnormal***
  - ↳ Virtual VAD Dataset
  - ↳ 3D 렌더링된 데이터 셋 (Virtual Scene + Character + Object)
    - ✓ 장면/캐릭터/물체 각각에 애니메이션을 넣을 수 있는 3D렌더러를 이용해 합성 데이터 제작.
    - ✓ 실제로 찍은 데이터셋에 비해 **아주 많은 양의 Abnormal 데이터**
    - ✓ Train : Normal + **Abnormal**
      Test : Normal + Abnormal

Dataset Name	Normal/Abnormal↓	Is Virtual?
CUHK Avenue	7.02	X
ShanghaiTech	17.57	X
NWPU Campus	21.47	X
UBnormal	1.66	O



*Acsintoae, Andra, et al. "Ubnormal: New benchmark for supervised open-set video anomaly detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

2. PFMF

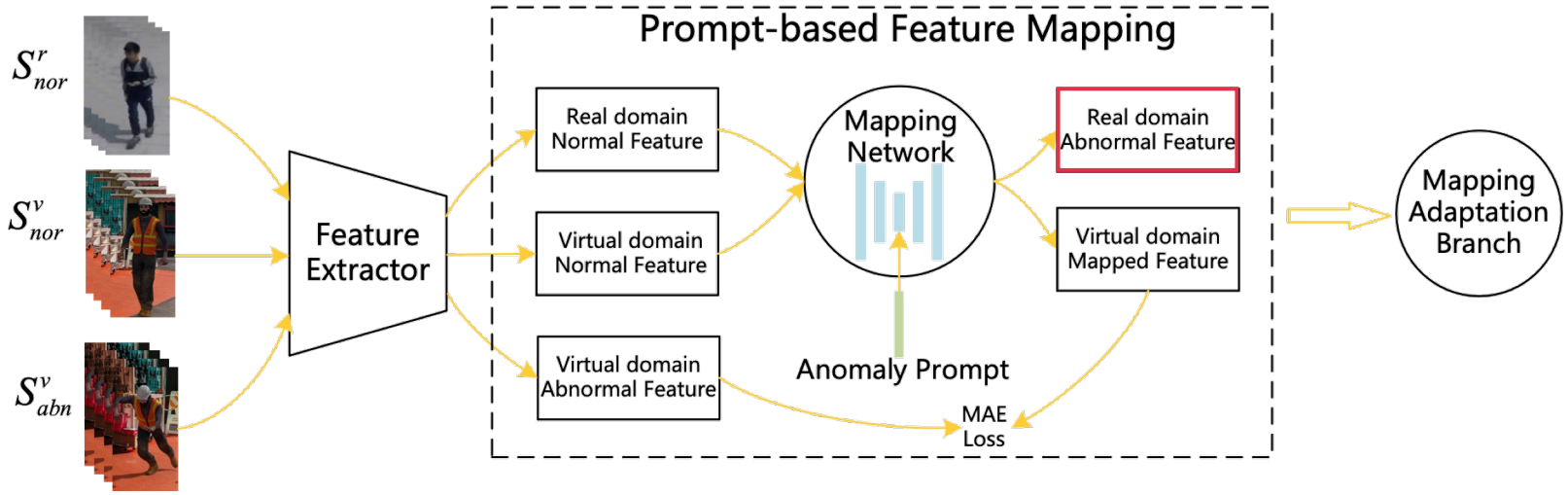
Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping

- **Issue of Virtual VAD Dataset**
  - ↳Issue #1 Anomaly Gap
    - ✓ 실제로 마주하는 데이터는 아주 방대한 Abnormal이 존재한다.
      BUT UBnormal(Virtual Data)에서는 한정된 분류를 가짐.
    - → 어떻게 방대한 Abnormal을 구현할 것인가?
- **Issue of VAD**
  - ↳Issue #2 Scene Gap
    - ✓ Scene-Specific Anomalies: 같은 사건에 대해 **환경에 따라 (ab)normal 여부**가 달라짐.
    - ✓ Scene-Specific Attributes: 카메라가 바라보는 **시점에 따라 성능에 영향**을 끼침.
      BUT UBnormal(Virtual Data)에서는 특정 시점과 환경을 가짐.
    - → 어떻게 Scene 환경과 시점을 극복하여 Abnormal을 감지할 것인가?

2. PFMF

Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping

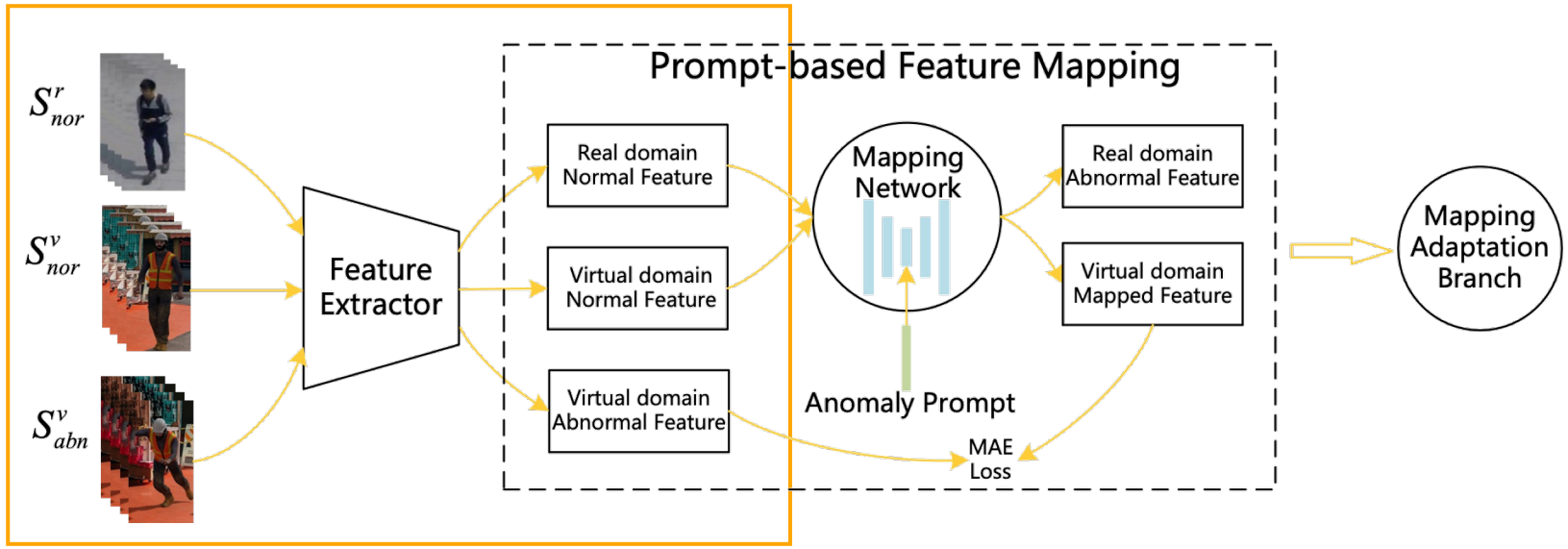
- **Method Overview – Real Dataset의 Abnormal을 보완하기**
  - ↳Virtual Dataset(가상 데이터셋) = Normal + Abnormal
  - ↳Real Dataset(현실 데이터셋) = Normal + ~~Abnormal~~
  - ↳Virtual Dataset을 이용해 Normal → Abnormal 매핑하는 Network를 구함.
    이 Net.으로 Real Dataset의 Normal → Abnormal을 구함.
    - ✓ 이 때 Generative Model(VAE)를 통해 얻은 Scene과 Anomaly Feature를 Anomaly Prompt를 추가로 조건을 넣음
  - ↳Virtual과 Real Dataset 간의 Domain Gap은 Mapping Adaptation Branch의 Domain Classifier를 통해 해결.



2. PFMF

Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping

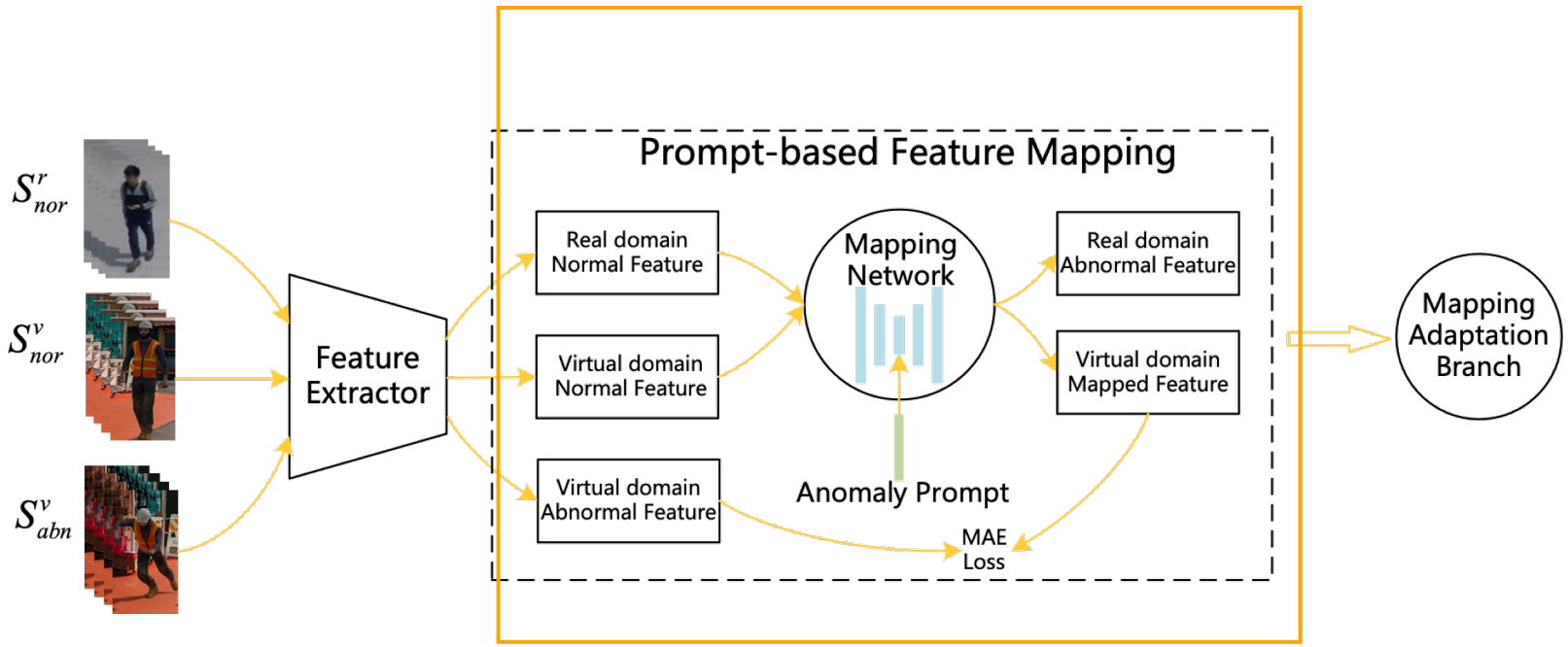
- **Method – Data PreProcessing and Feature Extract**
  - ↳Virtual Dataset의 Abnormal의 Bounding Box를 바로 얻을 수 있지만,
  - ↳Real Dataset의 Abnormal의 Bounding Box는 Annotation이 필요함.
    - ✓ YOLOv3를 통해 Bounding Box를 만들어준다.
    - ✓ Bounding Box를 이용해 Crop하여 Feature Extractor의 입력으로 넣는다.



2. PFMF

Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping

- **Method – Feature Mapping Network**



2. PFMF

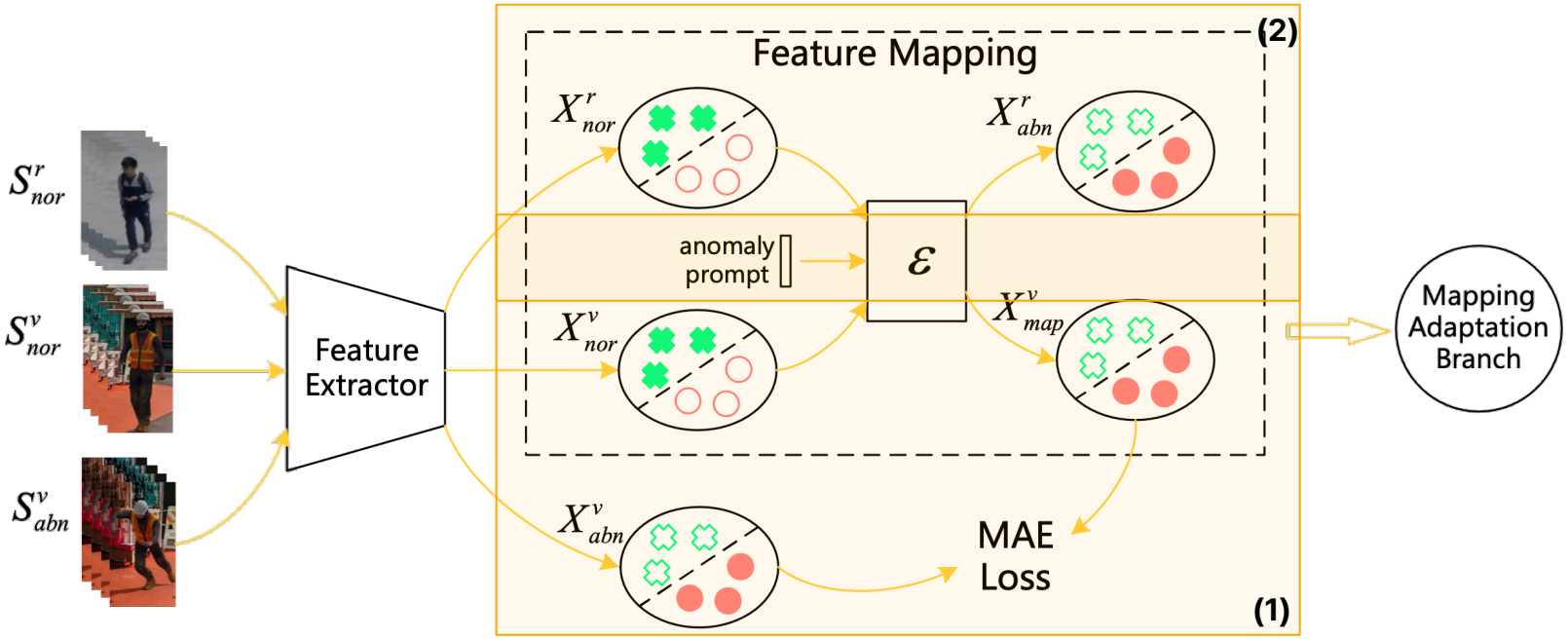
Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping

- **Method – Feature Mapping Network** [·]:Concat

↳Virtual로 학습된 Mapping Network를 통해 Real Dataset의 Normal → Abnormal을 구함.

(1) $X^v_{map} = \varepsilon_d([\varepsilon_e(X^v_{nor}), p_v]), \quad min\, MAE(X^v_{abn}, X^v_{map})$

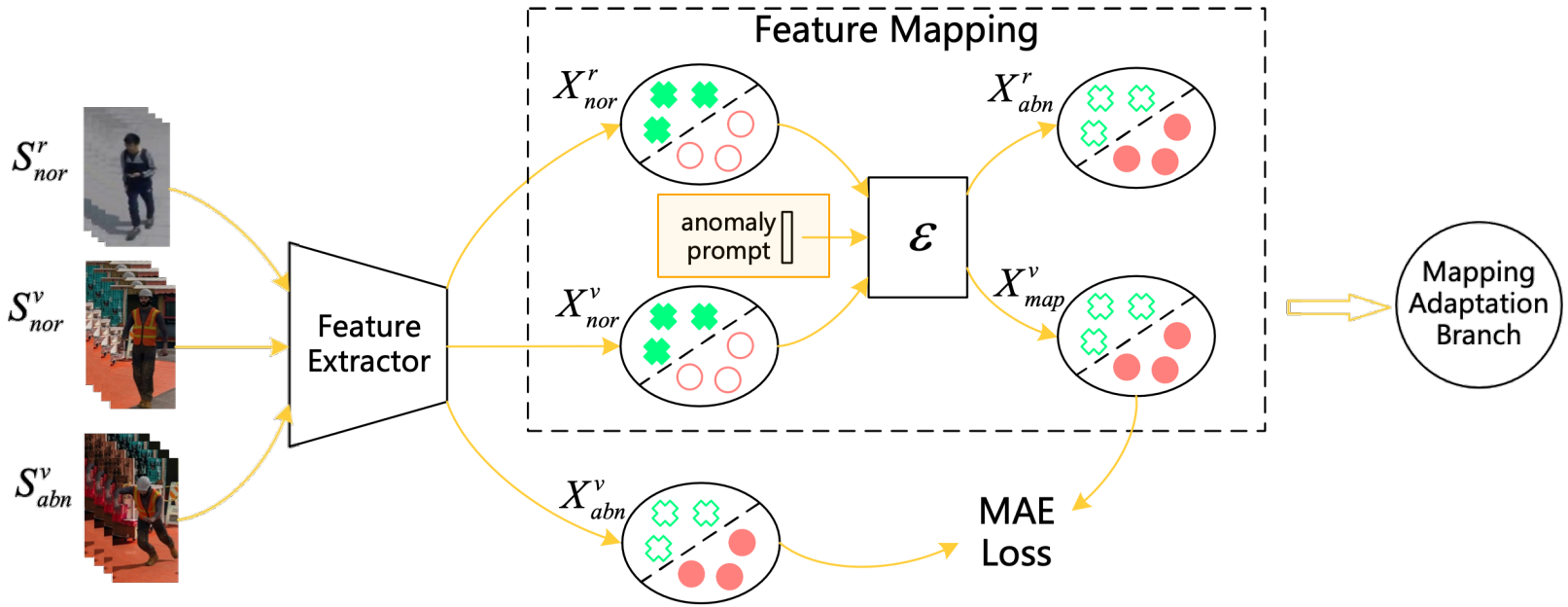
(2) $X^r_{abn} = \varepsilon_d([\varepsilon_e(X^r_{nor}), p_r]),$



2. PFMF

Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping

- **Method – Anomaly Prompt**



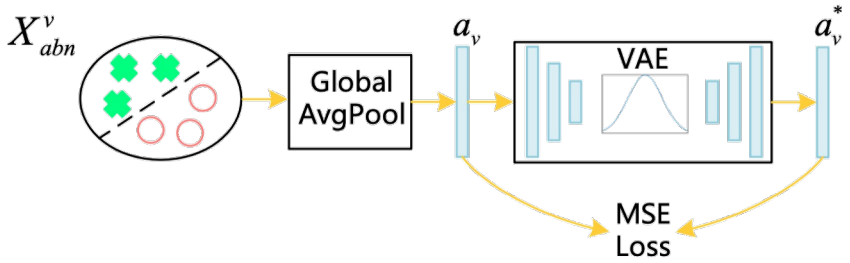
2. PFMF

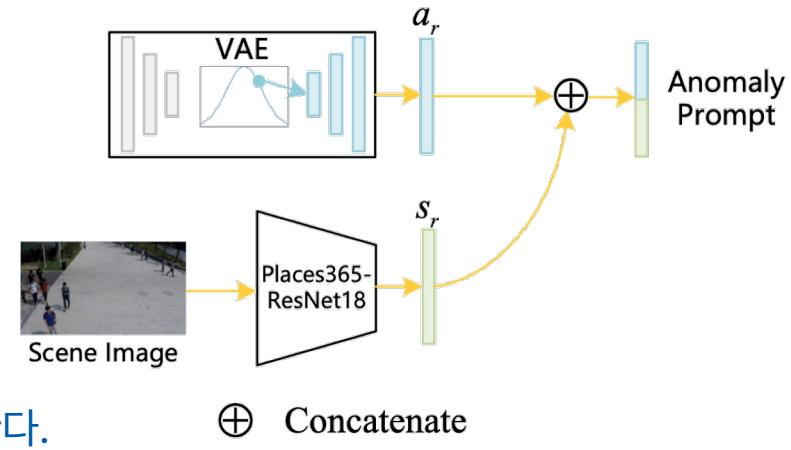
Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping

- **Method – Anormal Prompt : Feature Mapping을 위한 추가적인 Condition으로**

  = Anomaly Vector + Scene Vector

  ↳ Anomaly Vector : Virtual Abnormal Dataset으로 VAE를 학습함.





  ↳ Scene Vector : Place365 Pretrained ResNet18

  ✓ Real/Virtual abnormal feature에 대해 각각 Abnormal Prompt를 구한다.

  - Real Abnormal Prompt

  → Anomaly Vector : Real abnormal feature로 latent variable z를 구해 학습한 VAE에서 생성

  → Scene Vector : Place365 Pretrained ResNet18 with Real abnormal scene

  ↳ 어떤 Real Abnormal Data의 Feature가 와도 VAE를 이용해 대응할 수 있다.

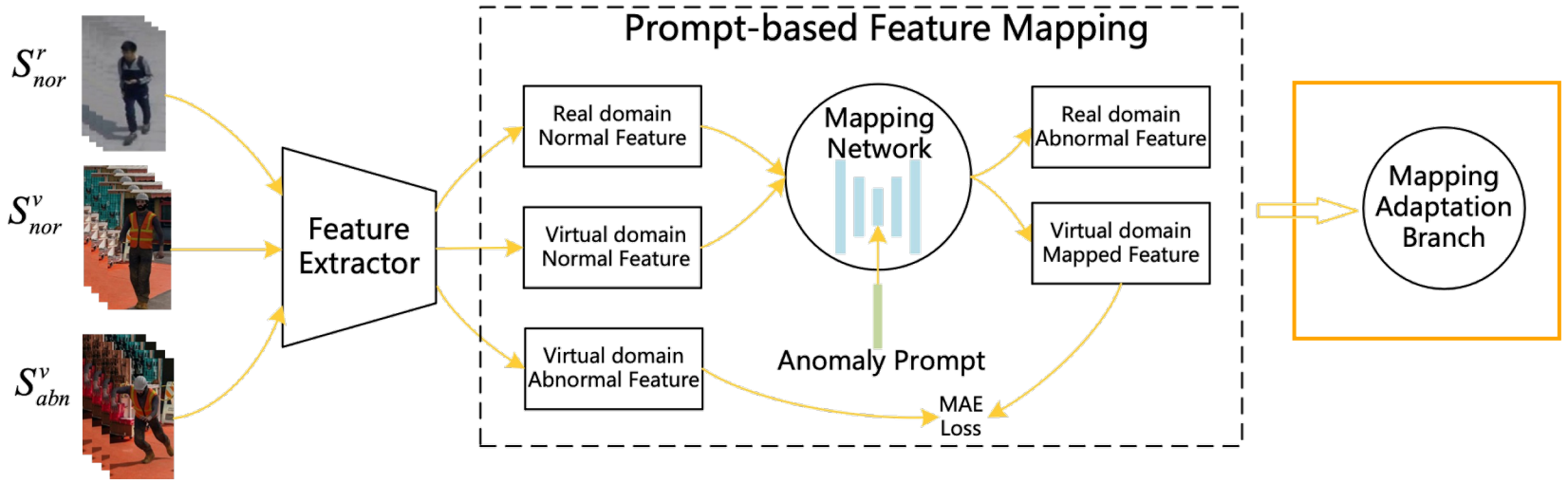
  ✓ Virtual 만 학습하였지만 Real 데이터가 와도 대응이 가능하다.

  ✓ Real과 Virtual의 Align은 이 후의 Mapping Adaptation Branch을 통해 가능함.

2. PFMF

Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping

- **Method – Mapping Adaptation Branch & Gradient Reversal Layer(GRL)**



2. PFMF

Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping

- **Method – Mapping Adaptation Branch & Gradient Reversal Layer(GRL)**

  ↳GRL은 pseudo function으로

  ✓ Forward-propagation: 입력 X를 그대로 전달

  ✓ Backward-propagation: 계산된 Gradient의 (-)값을 곱해 전달

$$R(X) = XI, \frac{dR}{dX} = -\lambda I$$

  ↳Anomaly Classifier

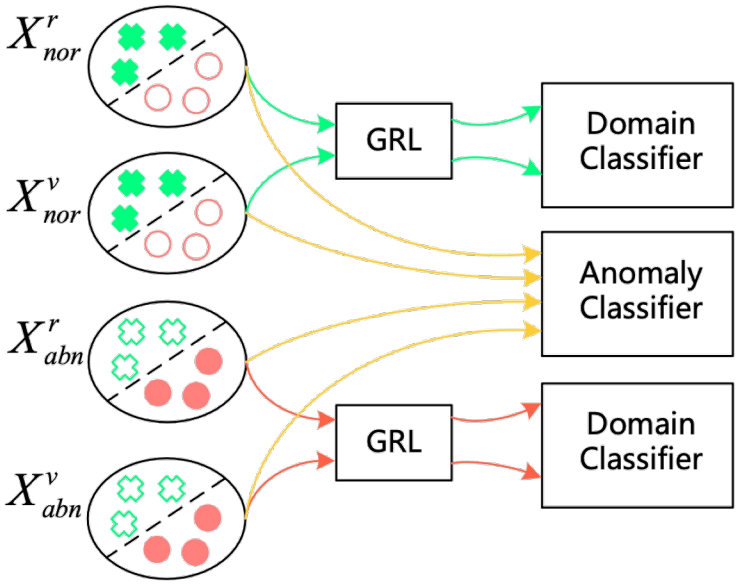
  ✓ 앞에서 Normal →Abnormal과정을 이용해 비슷한 Scene이라도 Abnormal이 있는 경우를 자연스럽게 가능하며, 이를 더 구별되게 해줌.

  ↳Domain Classification with GRL

  ✓ Why? Two Domain Classifier

  →Real/Virtual 간의 Domain Gap을 줄이기 위함.

  →이는 Real/Virtual 간의 공통된 특징을 찾기 위함임.

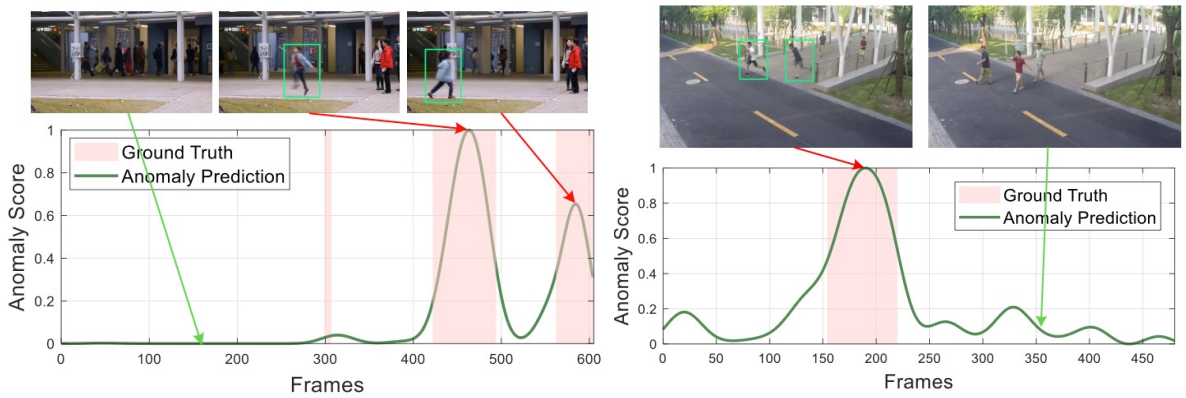


2. PFMF

Generating Anomalies for Video Anomaly Detection with Prompt-based Feature Mapping

- **Experiments**
- **ROC-AUC**
  - ↳ Macro ROC-AUC
    각 Class를 한 묶음으로 보았을 때의 **평균 ROC-AUC**
  - ↳ Micro ROC-AUC
    전체 Class를 한 묶음으로 보았을 때의 **ROC-AUC**

feature mapping	anomaly prompt	mapping adaptation	AUC	
			Micro	Macro
-	-	-	73.6	74.5
✔	-	-	78.9	84.2
✔	-	✔	80.9	86.5
✔	✔	-	80.0	85.3
✔	✔	✔	**83.8**	**87.8**



year	Method	Avenue AUC		ShanghaiTech AUC	
		Micro	Macro	Micro	Macro
2020	Park et al. [35]	88.5	-	70.5	-
	Sun et al. [43]	89.6	-	74.7	-
	Lu et al. [30]	85.8	-	77.9	-
	Wang et al. [48]	87.0	-	79.3	-
	Yu et al. [52]	89.6	-	74.8	-
	Tang et al. [44]	85.1	-	73.0	-
2021	Wang et al. [46]	88.3	-	76.6	-
	Astrid et al. [3]	87.1	-	73.7	-
	Liu et al. [28]	91.1	-	76.2	-
	Madan et al. [33]	88.6	-	74.6	-
	Li et al. [21]	88.8	-	73.9	-
	Georgescu et al. [11]	91.5	91.9	82.4	89.3
	Georgescu et al. [12]	92.3	90.4	82.7	89.3
2022	Zaheer et al. [54]	74.2	-	79.6	-
	Li et al. [23]	82.0	-	-	-
	Zaheer et al. [53]	-	-	69.9	-
	Cho et al. [5]	88.0	-	76.3	-
	Zhong et al. [55]	89.0	-	74.5	-
	Ristea et al. [39]*	92.9	91.9	83.6	89.5
	Acsintoae et al. [2]▽*	93.0	93.2	83.7	90.5
	PFMF (ours)▽	91.8	92.3	83.8	87.8
	PFMF (ours)▽*	**93.6**	**93.9**	**85.0**	**91.4**

▽ These methods apply virtual dataset for training.
* These methods use multi-task model ([11] or [12]) as backbone.

CVSP Lab
Computer Vision & Signal Processing

Ristea, Nicolae-C., et al.

Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors

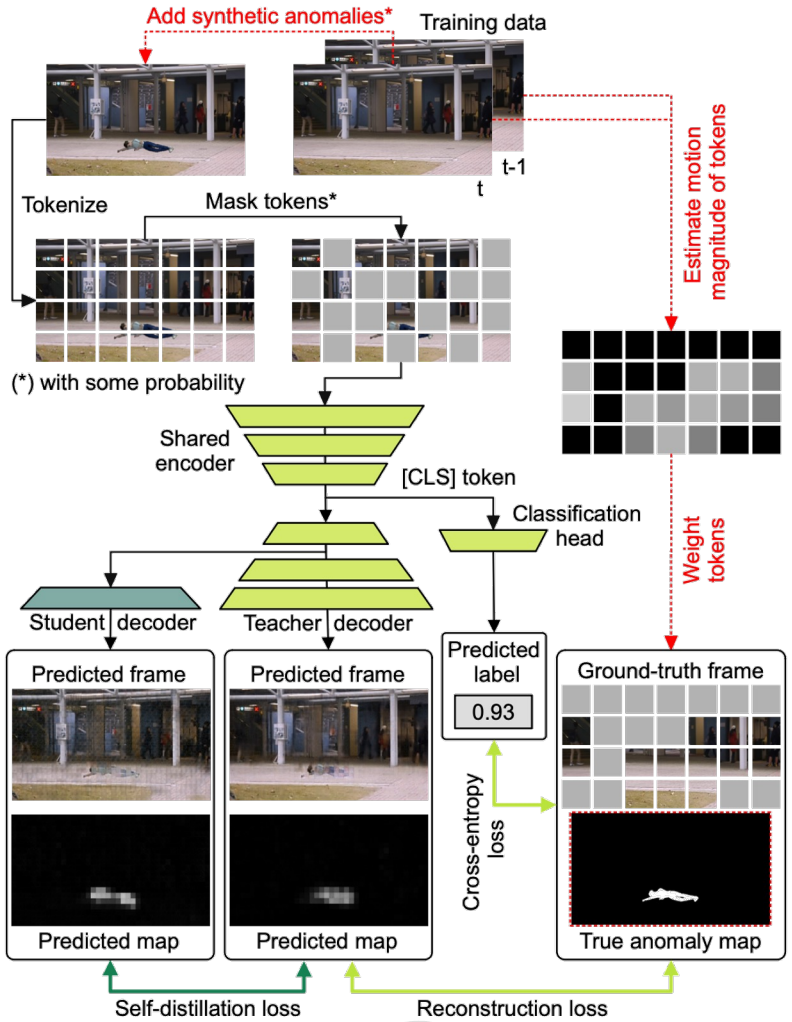
CVPR 2024

Ristea, Nicolae-C., et al. "Self-distilled masked auto-encoders are efficient video anomaly detectors." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

3. AED-MAE

Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors

- **State-of-the-Art(SotA) Model은 Object Detection Model에 의존적임**
  - ↳Object기반의 VAD Model은 SotA의 성능 지니지만 Object Detection Model이 필수이므로, **성능에 Dependancy**가 생김
    - ✓ 심지어, VAD Model보다 Object Detection에 더 긴 Inference time이 필요함.
    - ✓ 실시간성에 있어 적합하지 않음.
- **AED-MAE에서 Inference Time을 줄인 방법**
  - ↳크기를 많이 줄이고 더 효율적으로 연산함.
    - ✓ **Masked AutoEncoder**의 Visions Transformer(ViT)를 Convolutional vision Transformer(CvT)로 변경
    - ✓ **Layer 수를 아주 적게** 하였음.
    - ✓ **Self-Distillation**에서는 **Encoder를 공유**하도록함.
    - ✓ Motion Gradients를 이용해 장면의 **변화가 큰 것을 우선적으로 반영**함.



Ristea, Nicolae-C., et al. "Self-distilled masked auto-encoders are efficient video anomaly detectors." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

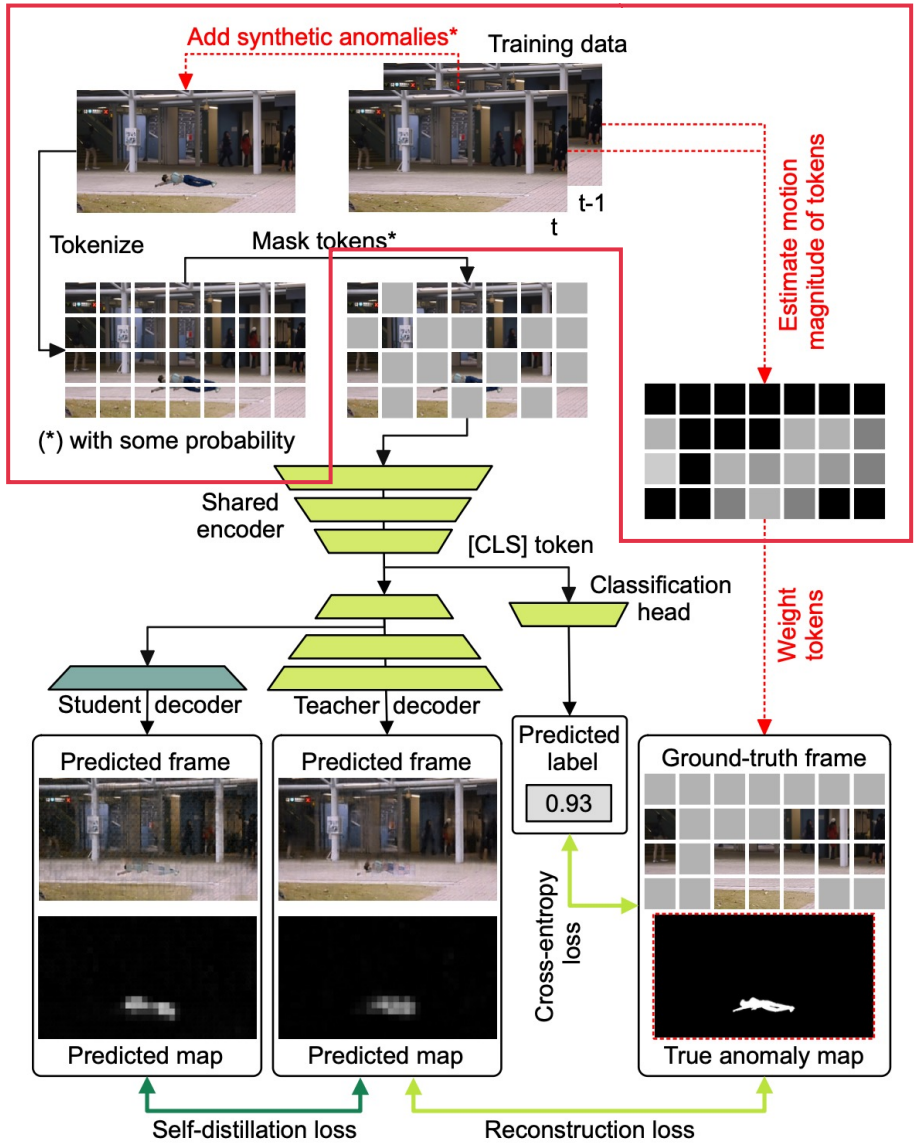
3. AED-MAE

Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors

- **Dataset Augmentation**
  - ↳ Temporal Consistancy만 만족하게 붙혀넣음
- **Method Get Motion Gradients**
  - ↳ 데이터 특성상 Fixed-View이므로 배경 자체보단 변화하는 부분을 더 집중시키도록 하기 위해 계산.
    - ✓ Patch의 Reconstruction/Self-distillation Loss에서 적용함.
  - ↳ 이전 프레임을 현재 프레임으로 빼준 것.
  - ↳ 패치의 max gradient가 해당 패치의 Motion Gradient 값이다.
  - ↳ 한 프레임에 대해 Motion Gradient를 min-max Normalization
    - ✓ 추가적으로 Synthetic 데이터에 대해서도 잘 적용하기 위해





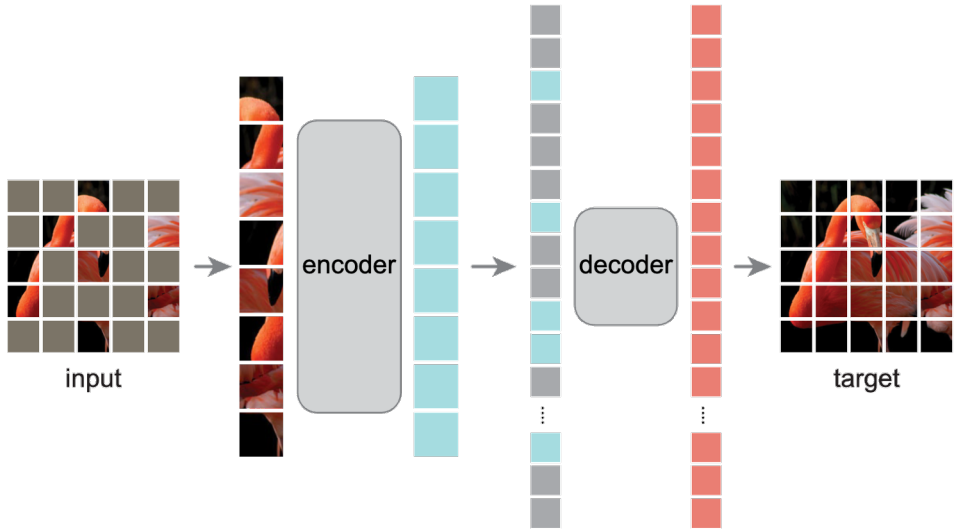


Ristea, Nicolae-C., et al. "Self-distilled masked auto-encoders are efficient video anomaly detectors." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

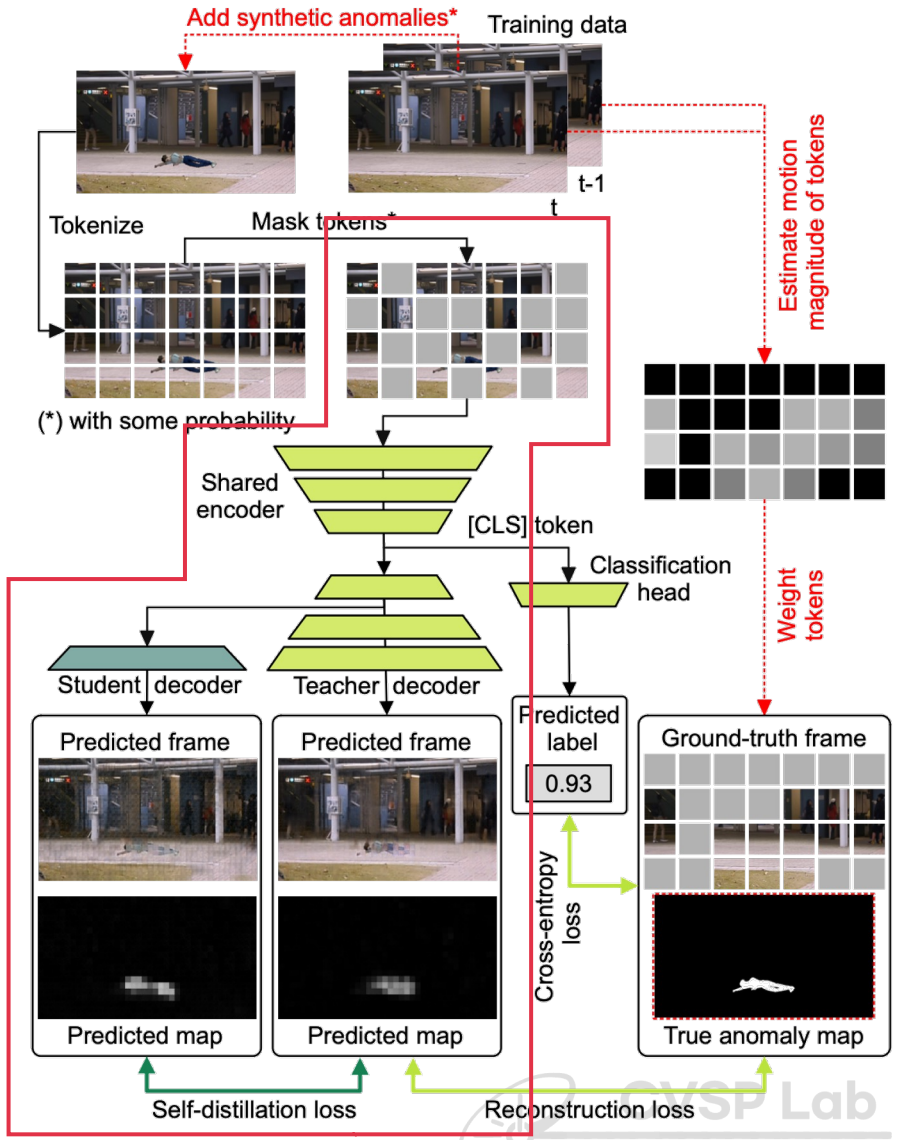
3. AED-MAE

Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors

- **MaskedAE**



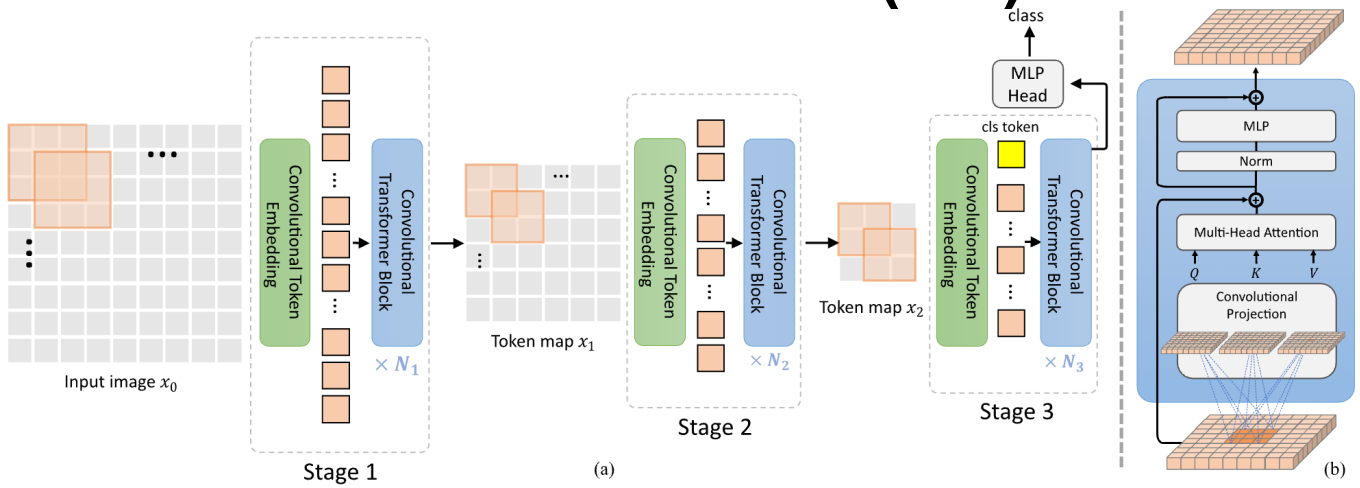
↳Patch(Token)로 나누어 Random Masking
↳Non-Masked Patch만 Feature Encoding
↳얻은 Patch의 Feature와 원본의 Patch 수를 맞추어 Decoding
↳이 때 Encoder/Decoder는 Vision Transformer



3. AED-MAE

Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors

- **Convolutional vision Transformer(CvT)**



↳Linear Projection이 사용되었던 것을 Convolutional Projection으로 변경

↳Postitional Encoding 안함.

[Table 3]
Wu, Haiping, et al. "Cvt: Introducing convolutions to vision transformers."
Proceedings of the IEEE/CVF international conference on computer vision. 2021.

Network	#Param. (M)	image size	FLOPs (G)	ImageNet top-1 (%)	Real top-1 (%)	V2 top-1 (%)
ResNet-50 [14]	25	224^2	4.1	76.2	82.5	63.3
ResNet-101 [14]	45	224^2	7.9	77.4	83.7	65.7
ResNet-152 [14]	60	224^2	11	78.3	84.1	67.0
ViT-B/16 [10]	86	384^2	55.5	77.9	83.6	–
ViT-L/16 [10]	307	384^2	191.1	76.5	82.2	–
DeiT-S [29][arxiv 2020]	22	224^2	4.6	79.8	85.7	68.5
DeiT-B [29][arxiv 2020]	86	224^2	17.6	81.8	86.7	71.5
Ours: CvT-13	20	224^2	4.5	81.6	86.7	70.4
Ours: CvT-21	32	224^2	7.1	82.5	87.2	71.3
Ours: CvT-13$_{\uparrow 384}$	20	384^2	16.3	83.0	87.9	71.9
Ours: CvT-21$_{\uparrow 384}$	32	384^2	24.9	**83.3**	**87.7**	**71.9**
BiT-M$_{\uparrow 480}$ [17]	928	480^2	837	85.4	–	–
ViT-B/16$_{\uparrow 384}$ [10]	86	384^2	55.5	84.0	88.4	–
ViT-L/16$_{\uparrow 384}$ [10]	307	384^2	191.1	85.2	88.4	–
ViT-H/16$_{\uparrow 384}$ [10]	632	384^2	–	85.1	88.7	–
Ours: CvT-13$_{\uparrow 384}$	20	384^2	16	83.3	88.7	72.9
Ours: CvT-21$_{\uparrow 384}$	32	384^2	25	84.9	89.8	75.6
Ours: CvT-W24$_{\uparrow 384}$	277	384^2	193.2	**87.7**	**90.6**	**78.8**

Ristea, Nicolae-C., et al. "Self-distilled masked auto-encoders are efficient video anomaly detectors." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

3. AED-MAE

Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors

- **MaskedAE with CvT**

Normalized patch motion gradient

$$\mathcal{L}_{\text{wMSE}}(\boldsymbol{x}_t, \theta_T) = \frac{1}{n}\sum_{i=1}^{n} w_i^{(t)} \cdot \|\boldsymbol{p}_i^{(t)} - \hat{\boldsymbol{p}}_i^{(t)}\|_2^2,$$

↳인코더/디코더에 각 3-Layer 4-Head CvT

- **Self-Distillation**

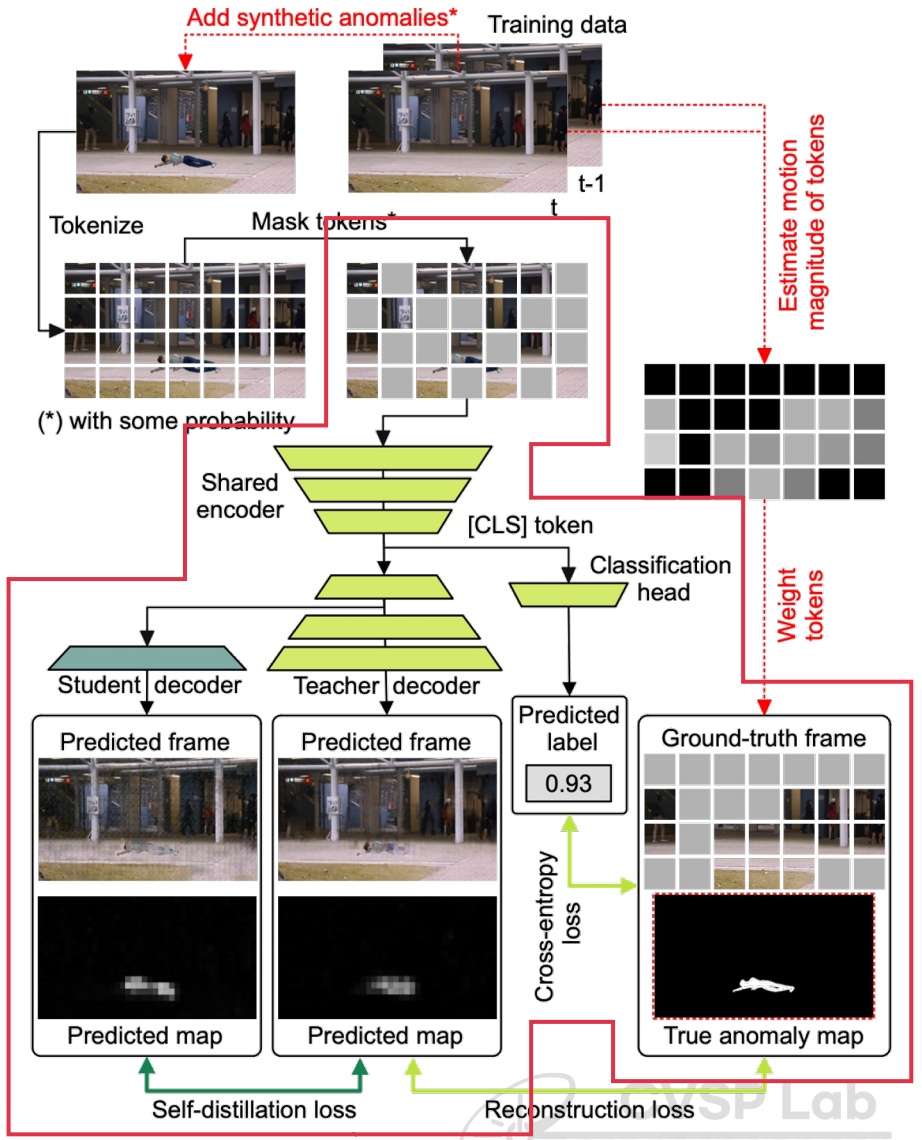
$$\mathcal{L}_{\text{SD}}(\hat{\boldsymbol{x}}_t, \theta_S) = \frac{1}{n}\sum_{i=1}^{n} w_i^{(t)} \cdot \|\hat{\boldsymbol{p}}_i^{(t)} - \tilde{\boldsymbol{p}}_i^{(t)}\|_2^2,$$

↳Teacher와 Encoder를 공유함.

↳Teacher의 첫 번째 Decoder를 지나고 Student Decoder 학습

- **Classification Head**

↳MAE w. CvT Decoder로 나온 Feature를 이용해 측정



부산대학교 PUSAN NATIONAL UNIVERSITY
CVSP Lab Computer Vision & Signal Processing

3. AED-MAE

Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors

- **Ablation Study**
  - ↳각각의 모듈이 추가될 수록 성능이 좋아지는 추세임.
  - ↳Self-Distillation → Micro AUC 대폭 증가
  - ↳Synthetic Data 사용 = Anomaly maps, Classification Head들을 Supervised Learning가능
    → Micro/Macro AUC 대폭 증가

Motion weights	Self-distillation	Synthetic data	Anomaly maps	Classif. head	Avenue AUC		Shanghai AUC	
					Micro	Macro	Micro	Macro
					84.0	85.6	69.7	80.1
✓					84.8	86.3	71.3	80.9
✓	✓				88.5	86.0	76.3	83.8
✓	✓	✓			88.5	86.9	77.0	83.0
✓	✓	✓	✓		90.5	89.6	77.3	84.2
✓	✓	✓	✓	✓	**91.3**	**90.9**	**79.1**	**84.7**

Data set	Measure	Percentage of synthetic data			
		0%	25%	50%	75%
Avenue	Micro AUC	88.5	**91.3**	90.6	89.9
	Macro AUC	86.0	**90.9**	89.4	87.7
Shanghai	Micro AUC	76.3	**79.1**	77.9	77.7
	Macro AUC	83.8	**84.7**	84.4	84.1

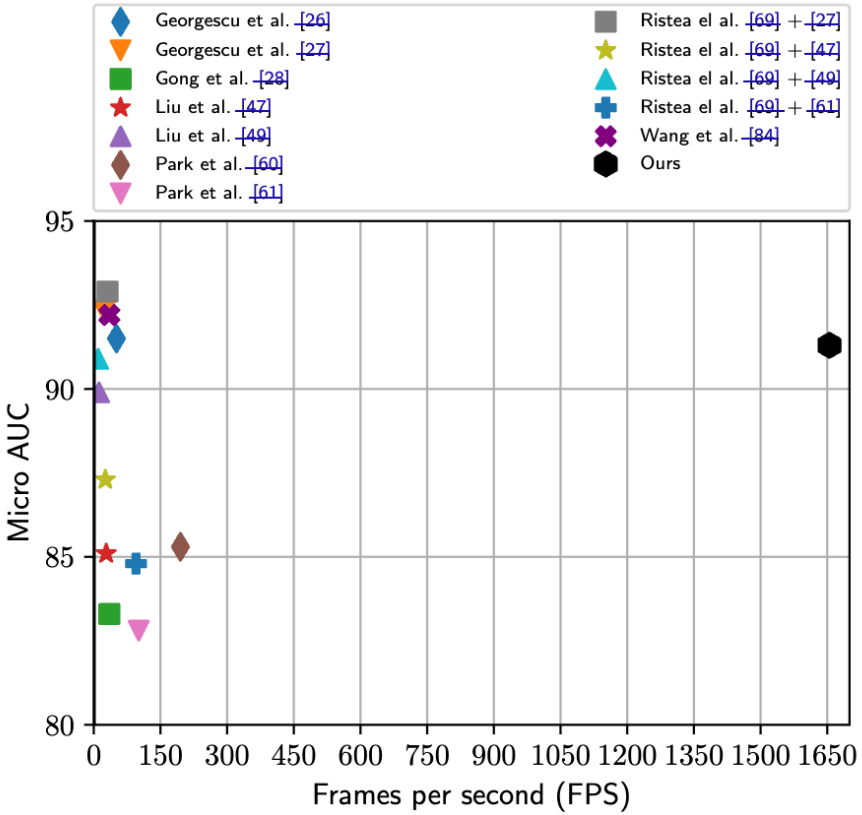
Ristea, Nicolae-C., et al. "Self-distilled masked auto-encoders are efficient video anomaly detectors." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

3. AED-MAE

Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors

- **Experiments**

  ↳압도적 실시간 성능 그렇다고 성능이 떨어지는 것도 아님.



Method	GFLOPs ↓	#Params (M) ↓	FPS ↑
Georgescu *et al.* [26]	107.9	65	51
Georgescu *et al.* [27]	121.6	67	24
Liu *et al.* [49]	179.5	320	12
Park *et al.* [60]	84	64	195
Gong *et al.* [28]	55.2	6	35
Ours	**0.8**	**3**	**1655**

@RTX 3090

Type	Method	Avenue AUC Micro	Avenue AUC Macro	Shanghai AUC Micro	Shanghai AUC Macro	UBnormal AUC Micro	UBnormal AUC Macro	Ped2 AUC Micro	Ped2 AUC Macro	FPS
Frame or cube level	[4]	87.1	-	75.9	-	-	-	96.5	-	-
	[5]	84.7	-	73.7	-	-	-	98.4	-	-
	[9]	-	-	-	-	68.5	80.3	-	-	37
	[28]	83.3	-	71.2	-	-	-	94.1	-	35
	[37]	88.9	-	-	-	-	-	-	-	-
	[43]	90.0	-	-	-	-	-	96.6	-	-
	[47]	85.1	81.7	72.8	80.6	-	-	95.4	-	28
	[48]	84.4	-	-	-	-	-	87.5	-	-
	[54] + [47]	89.1	84.8	74.6	83.3	-	-	-	-	26
	[54] + [61]	86.4	86.3	70.6	80.3	-	-	-	-	94
	[57]	86.9	-	-	-	-	-	96.2	-	-
	[60]	85.3	-	72.2	-	-	-	96.3	-	195
	[61]	82.8	86.8	68.3	79.7	-	-	97.0	-	101
	[62]	72.0	-	-	-	-	-	88.3	-	-
	[63]	87.2	-	-	-	-	-	93.0	-	-
	[66]	-	-	-	-	-	-	93.5	-	-
	[67]	-	-	-	-	-	-	88.4	-	-
	[69] + [47]	87.3	84.5	74.5	82.9	-	-	-	-	26
	[69] + [61]	84.8	88.6	69.8	80.2	-	-	-	-	95
	[76]	84.6	-	-	-	-	-	-	-	-
	[77]	-	-	-	76.5	50.3	76.8	-	-	56
	[78]	89.6	-	74.7	-	-	-	-	-	-
	[80]	85.1	-	73.0	-	-	-	96.3	-	-
	[82]	-	-	76.1	-	-	-	-	-	-
	[87]	87.0	-	79.3	-	-	-	-	-	-
	[89]	-	-	80.4	-	-	-	-	-	-
	[90]	86.6	-	-	-	-	-	96.9	-	-
	[93]	90.1	-	78.6	-	62.7	-	-	-	-
	[96]	90.2	-	-	-	-	-	97.3	-	-
	[100]	-	-	78.9	-	-	-	-	-	-
	[103]	-	-	-	-	-	-	91.0	-	-
	Ours	91.3	90.9	79.1	84.7	58.5	81.4	95.4	98.4	1655

Computer Vision & Signal Processing

3. AED-MAE

Self-Distilled Masked Auto-Encoders are Efficient Video Anomaly Detectors

- **Experiments**

