Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection

CVPR 2022

Ristea Nicolae-Cătălin, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah.

Reporter: Yu-Chen Lai

Date: 2022/09/16

Outline

- Anomaly detection
- Proposed Method
  - Masked convolution
  - Channel attention module
- Experiments and Results
  - Dataset
  - Evaluation Metrics
  - Implementation Choices and Tuning
  - Result
    - Preliminary Experiments
    - Ablation Study
    - Anomaly Detection in Images
    - Abnormal Event Detection in Video
- Conclusion
- Disscusion

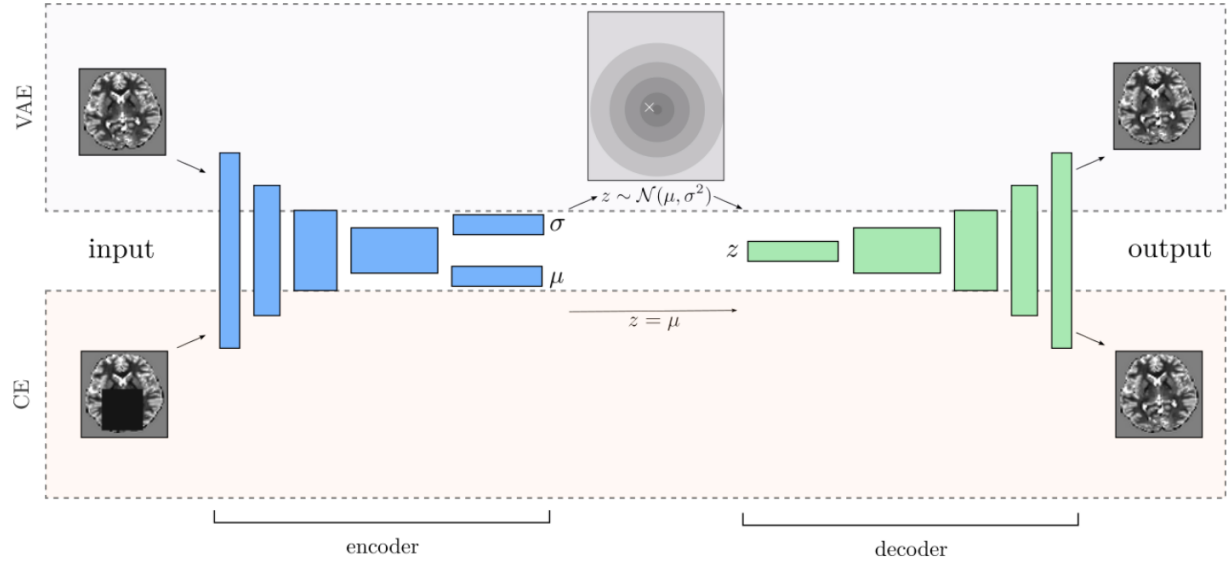
Anomaly detection

- Unsupervised - only train on **normal samples**
- The Anomaly (abnormal) samples - determined by the training set
- Categories
  - Dictionary learning
  - Distance-based
  - Probability-based
  - Change detection frameworks (Video)
  - **Reconstruction-based**

Zimmerer, David, Simon AA Kohl, Jens Petersen, Fabian Isensee, and Klaus H. Maier-Hein. "Context-encoding variational autoencoder for unsupervised anomaly detection." *arXiv preprint arXiv:1812.05941* (2018).

Introduction (1 /3)

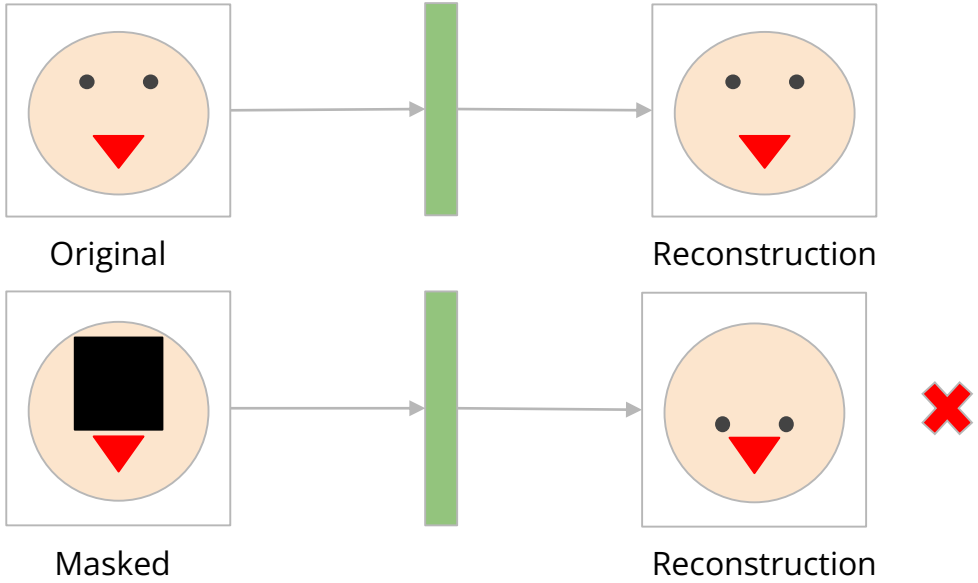
- A distinguished subcategory of reconstruction methods relies on **predicting masked information**.



Zimmerer, David, Simon AA Kohl, Jens Petersen, Fabian Isensee, and Klaus H. Maier-Hein. "Context-encoding variational autoencoder for unsupervised anomaly detection." *arXiv preprint arXiv:1812.05941* (2018).

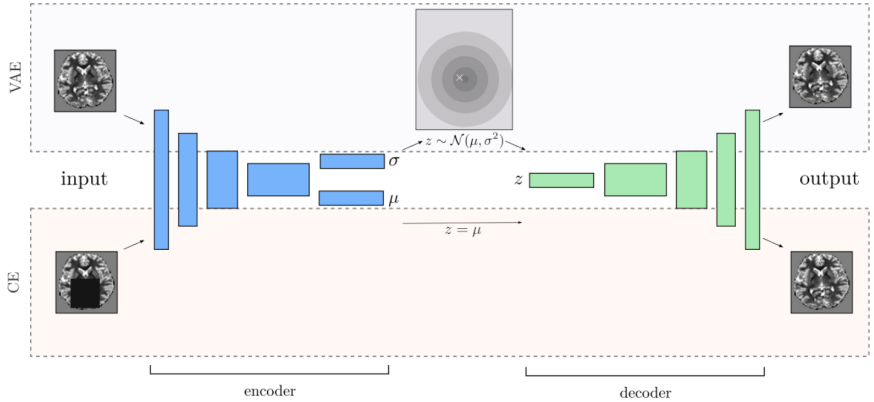
Introduction (2 /3)

- Why we mask information ?



Introduction (3 /3)

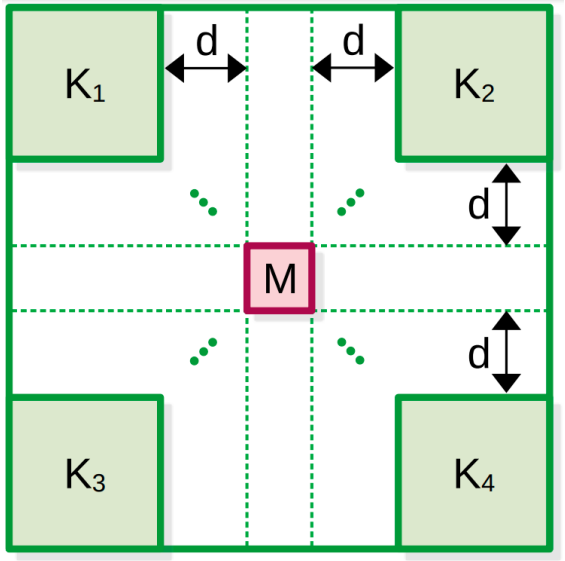
- **SSPCAB** integrates the capability of reconstructing the masked information into a **neural block**.
- Advantages
  - Mask information **at any layer** in a neural network (not only at the input)
  - Can be **integrated** into a wide range of neural architectures.



Zimmerer, David, Simon AA Kohl, Jens Petersen, Fabian Isensee, and Klaus H. Maier-Hein. "Context-encoding variational autoencoder for unsupervised anomaly detection." *arXiv preprint arXiv:1812.05941* (2018).

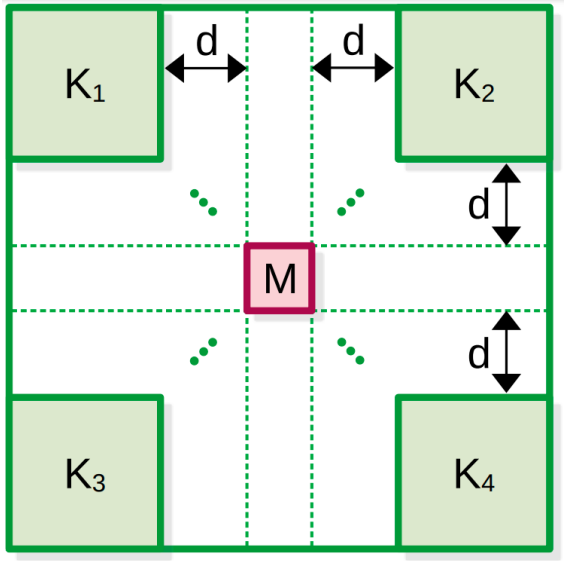
Masked convolution (1/2)

- The receptive field of convolutional filter
  - **Sub-kernels** $\boldsymbol{K}_i \in \mathbb{R}^{k' \times k' \times c}, \forall i \in \{1, 2, 3, 4\}$, where $k' \in \mathbb{N}^+$
  - Distance (dilation rate): $d \in \mathbb{N}^+$
  - The center of receptive field: $\boldsymbol{M} \in \mathbb{R}^{1 \times 1 \times c}$
  - Spatial size k of receptive field: $k = 2k' + 2d + 1$



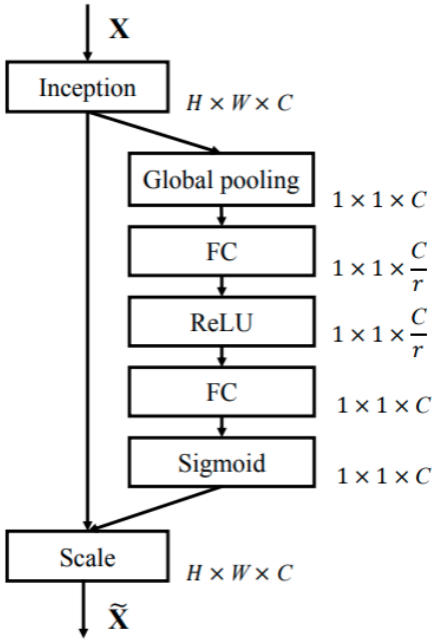
Masked convolution (2/2)

- Convolution operations
  - Zero-padding:$k' + d$ pixels around the input
  - Stride: 1
  - Number of c masked convolutional filters
  - Output tensor Z is passed through a ReLU activation
- The work of mask convolution
  - **Mask** information
  - **Reconstructio**n



Channel attention module

- Reduce $\boldsymbol{Z}$ to a vector $\boldsymbol{z} \in \mathbb{R}^c$ through a **global average pooling** performed on each channel.
- **Scale factors** $\boldsymbol{s} \in \mathbb{R}^c$ is computed as follows:
  $\boldsymbol{s} = \sigma\left(\boldsymbol{W}_2 \cdot \delta\left(\boldsymbol{W}_1 \cdot \boldsymbol{z}\right)\right)$
- where σ is the sigmoid activation
- δ is the ReLU activation
- $\boldsymbol{W}_1 \in \mathbb{R}^{\frac{c}{r} \times c}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{c \times \frac{c}{r}}$
- r is the reduction ratio



SE-Inception Module

Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141. 2018.

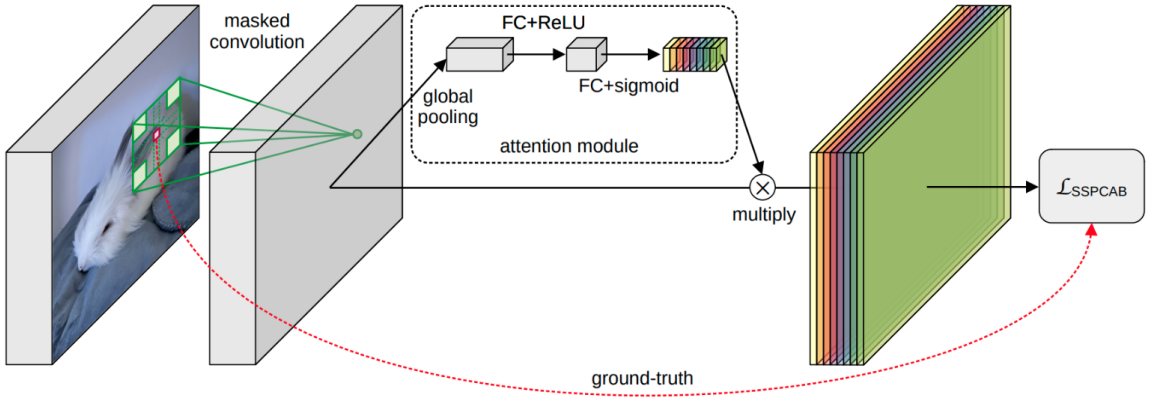
Reconstruction loss

- MSE between the input and the output

$$\mathcal{L}_{\text{SSPCAB}}(G, \boldsymbol{X}) = (G(\boldsymbol{X}) - \boldsymbol{X})^2 = \left(\hat{\boldsymbol{X}} - \boldsymbol{X}\right)^2$$

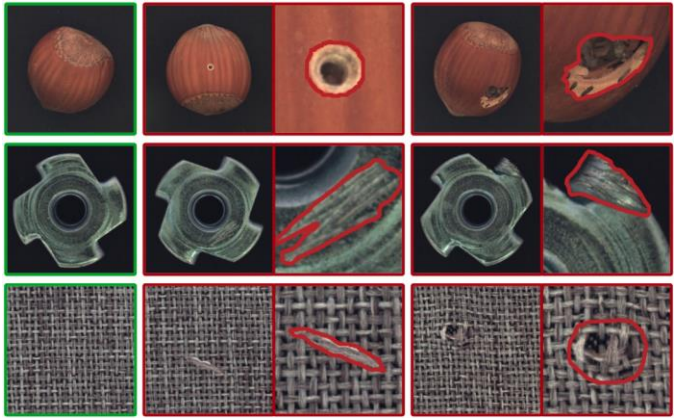
- When integrating SSPCAB into a neural model F having its own loss function $\mathcal{L}_F$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_F + \lambda \cdot \mathcal{L}_{\text{SSPCAB}}$$



Dataset (1/3)

- MVTec AD
  - A standard benchmark for evaluating AD methods on industrial inspection images.
  - It contains images from 10 object categories and 5 texture categories.
  - Defect-free training images - 3629
  - Test images with or without anomalies - 1725



Dataset (2/3)

- CHUK Avenue
  - A popular benchmark for video anomaly detection
  - Training videos - 16
  - Test videos - 21
  - The anomalies - people throwing papers, running, dancing, loitering, and walking in the wrong direction

Strange action	Wrong direction	Abnormal object
		

Dataset (3/3)

- ShanghaiTech
  - Training videos - 330
  - Test videos - 107
  - The anomalies - people fighting, stealing, chasing, jumping, and riding bike or skating in pedestrian zones

Evaluation Metrics (1/2)

- Image anomaly detection (On MVTec AD)
  - Area under the ROC curve (**AUROC**)
    - Detection task
    - TPR : the percentage of anomalous images that are correctly classified
    - FPR : the percentage of normal images mistakenly classified as anomalous
  - Average precision (**AP**)
    - Localization (segmentation) task
    - TPR : the percentage of abnormal pixels that are correctly classified
    - FPR : the percentage of normal pixels wrongly classified as anomalous

Bergmann, Paul, Michael Fauser, David Sattlegger, and Carsten Steger. "MVTec AD--A comprehensive real-world dataset for unsupervised anomaly detection." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9592-9600. 2019.

Evaluation Metrics (2/2)

- Video anomaly detection
  - Marking a frame as abnormal if at least one pixel inside the frame is abnormal.
  - **micro AUC**: computed after concatenating all frames from the entire test set
  - **macro AUC**: the average of the AUC scores on individual videos
  - Region-based detection criterion (**RBDC**)
    - RBDC takes each **detected region** into consideration, marking a detected region as true positive if the IoU with the ground-truth region is greater than a threshold α. We set $\alpha = 0.1$.
  - Track-based detection criterion (**TBDC**)
    - TBDC measures whether abnormal regions are accurately **tracked across time**. It considers a detected track as true positive if the number of detections in a track is greater than a threshold β. We set $\beta = 0.1$.

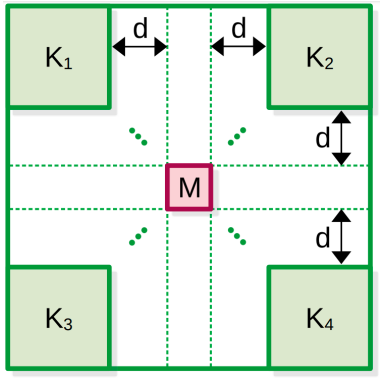
Implementation Choices and Tuning

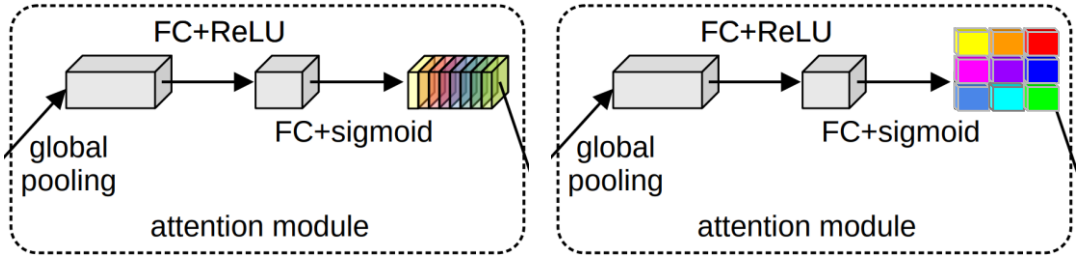
- Choose the underlying models for SSPCAB
- Replace the **penultimate convolutional layer** with SSPCAB in all underlying models.
- In a set of preliminary trials with a basic auto-encoder on Avenue, we tuned the hyperparameter λ, considering values between 0.1 and 1, at a step of 0.1.
- Decided to use **λ = 0.1**.

- Loss: $\mathcal{L}_{\text{total}} = \mathcal{L}_F + \lambda \cdot \mathcal{L}_{\text{SSPCAB}}$

Result 1 : Preliminary Experiments (1/4)

- Dataset: Avenue
- Hyperparameters of our masked convolution
  - k' : {1, 2, 3}
  - d : {0, 1, 2}
- Two alternative loss functions
  - Mean Absolute Error (**MAE**)
  - Mean Squared Error (**MSE**)
- Several types of attention (added after the masked convolution)
  - Channel attention (**CA**)
  - Spatial attention (**SA**)

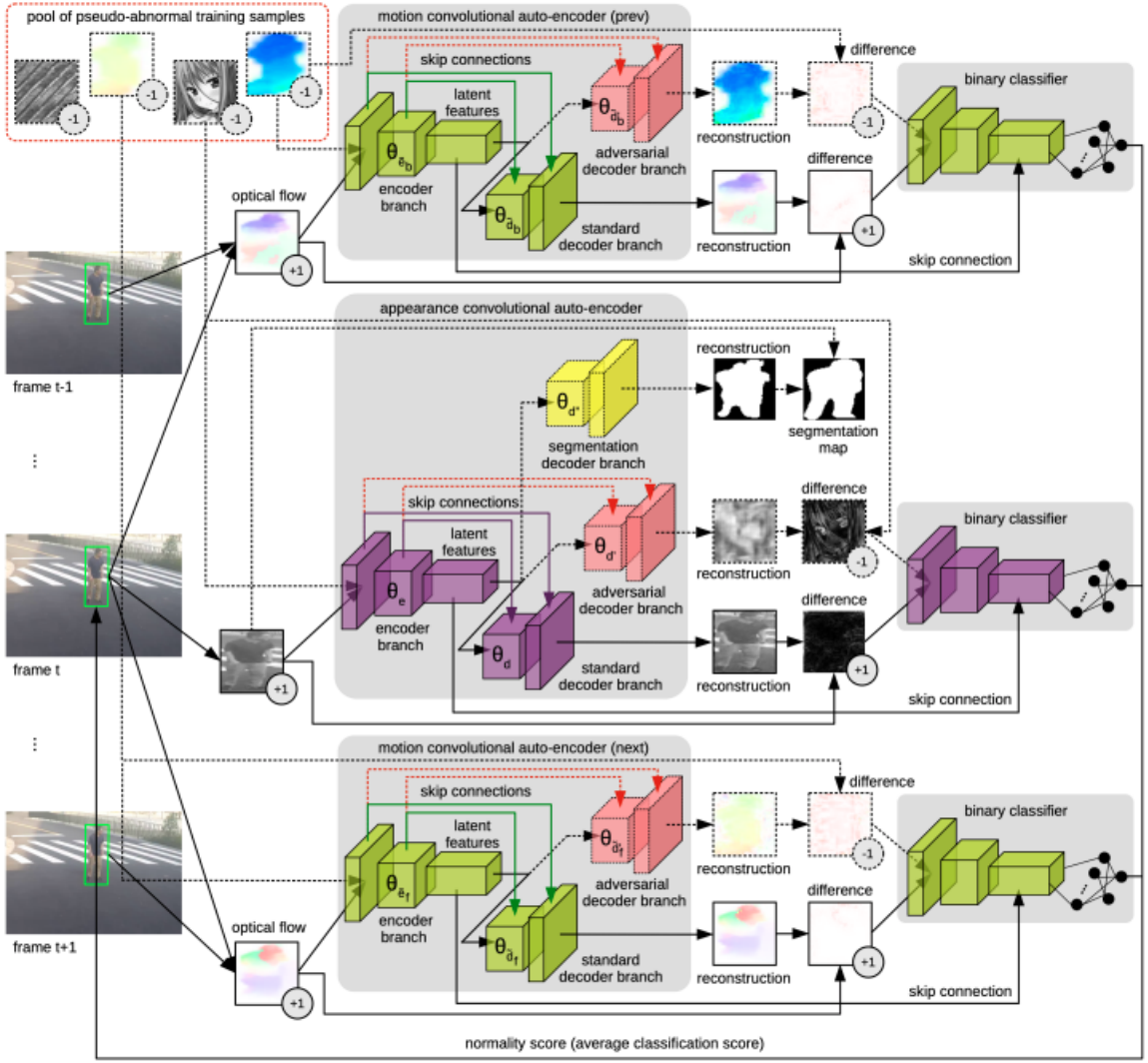




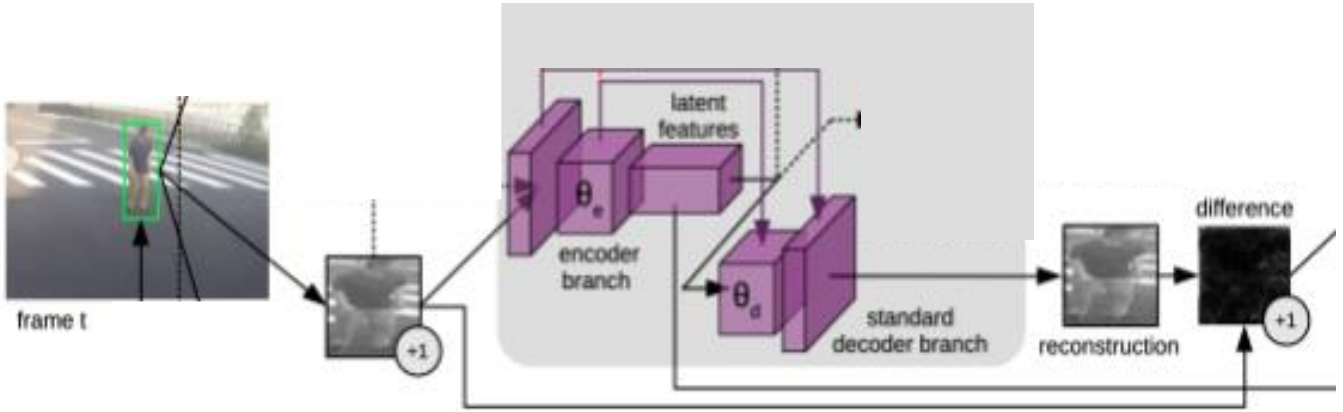
Result 1 : Preliminary Experiments (2/4)

- Baseline: The appearance convolutional auto-encoder from the paper.
- Stripping out the additional components such as optical flow, skip connections, adversarial training, mask reconstruction and binary classifiers.
- Only a **plain auto-encoder** in our preliminary experiments

Georgescu, Mariana Iuliana, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. "A background-agnostic framework with adversarial training for abnormal event detection in video." IEEE Transactions on Pattern Analysis and Machine Intelligence 44, no. 9 (2021): 4505-4523.


pool of pseudo-abnormal training samples
motion convolutional auto-encoder (prev)
skip connections
latent features
adversarial decoder branch
encoder branch
standard decoder branch
reconstruction
difference
binary classifier
optical flow
skip connection
appearance convolutional auto-encoder
segmentation decoder branch
segmentation map
frame t-1
frame t
frame t+1
motion convolutional auto-encoder (next)
normality score (average classification score)

Result 1 : Preliminary Experiments (3/4)



Georgescu, Mariana Iuliana, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. "A background-agnostic framework with adversarial training for abnormal event detection in video." IEEE Transactions on Pattern Analysis and Machine Intelligence 44, no. 9 (2021): 4505-4523.

Result 1 : Preliminary Experiments (4/4)

- Loss: MSE
- Sub-kernel size: 1
- Dilation rate: 1
- Reduction rate: 8

Method	Loss type	d	k'	r	Attention type	AUC		RBDC	TBDC
						Micro	Macro		
Plain auto-encoder	-	-	-	-	-	80.0	83.4	49.98	51.69
	MAE	0	1	-	-	83.3	84.1	47.46	52.11
		1	1	-	-	83.9	84.6	49.05	52.21
		2	1	-	-	83.2	84.3	48.56	52.03
	MSE	0	1	-	-	83.6	84.2	47.86	52.21
		1	1	-	-	84.2	84.9	49.22	52.29
		2	1	-	-	83.6	84.3	48.44	51.98
	MSE	0	2	-	-	83.7	84.0	47.41	53.02
		1	2	-	-	84.0	85.1	48.22	51.84
		2	2	-	-	82.7	83.1	46.94	50.22
	MSE	0	3	-	-	82.6	83.7	48.28	51.91
		1	3	-	-	82.9	84.7	48.13	52.07
		2	3	-	-	83.1	83.8	47.13	49.96
	MSE	1	1	8	CA	**85.9**	**85.6**	53.81	**56.33**
		1	1	-	SA	84.3	84.4	53.31	53.41
		1	1	8	CA+SA	85.7	85.6	**53.98**	54.11
	MSE	1	1	4	CA	85.6	85.3	53.83	55.99
		1	1	16	CA	84.4	84.9	53.28	54.37

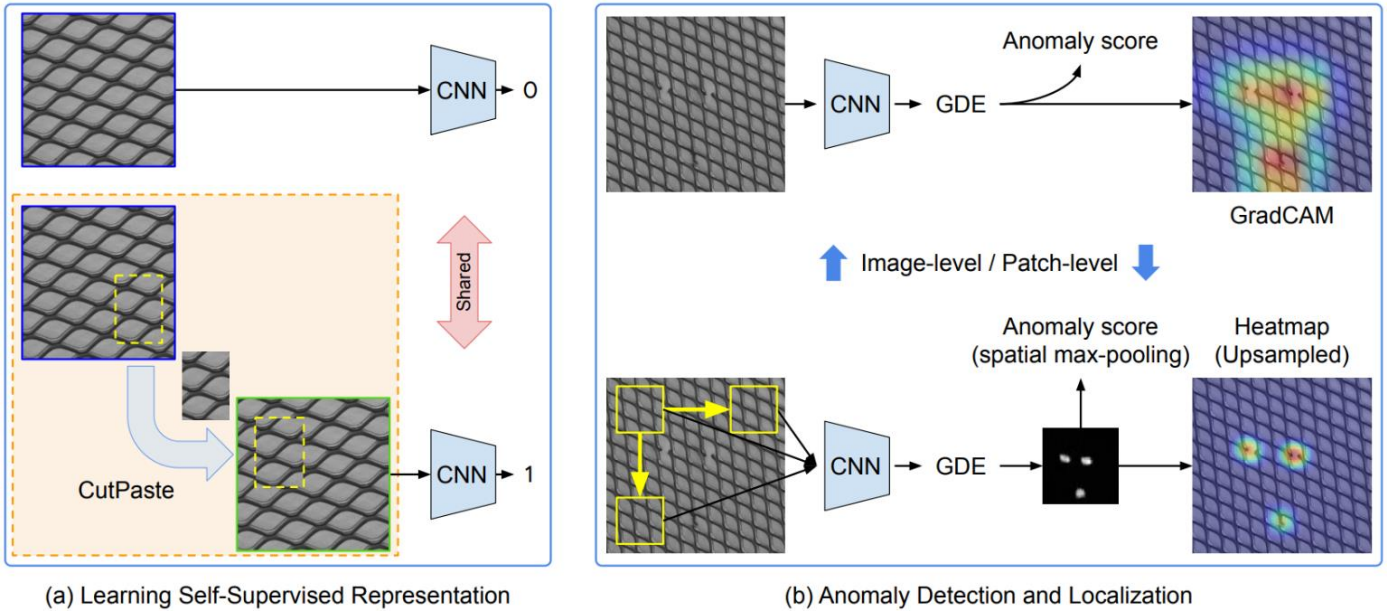
Result 2 : Ablation study

	Location of SSPCAB			AUC		RBDC	TBDC
	Early	Middle	Late	Micro	Macro		
Plain auto-encoder				80.0	83.4	49.98	51.69
	✓			81.1	83.6	50.86	52.44
		✓		84.2	85.0	52.73	54.02
			✓	85.9	85.6	53.81	56.33
	✓	✓		82.7	83.8	50.54	52.70
	✓		✓	83.2	84.1	52.33	53.01
		✓	✓	**86.1**	**85.7**	**54.03**	56.07
	✓	✓	✓	85.3	85.4	53.11	**56.64**

Size of M	AUC		RBDC	TBDC
	Micro	Macro		
	80.0	83.4	49.98	51.69
1×1	85.9	85.6	53.81	56.33
3×3	85.9	85.5	53.93	56.31

Result 3 : Anomaly Detection in Images (1/4)

- Baselines: **CutPaste [34]** and DRAEM [79]



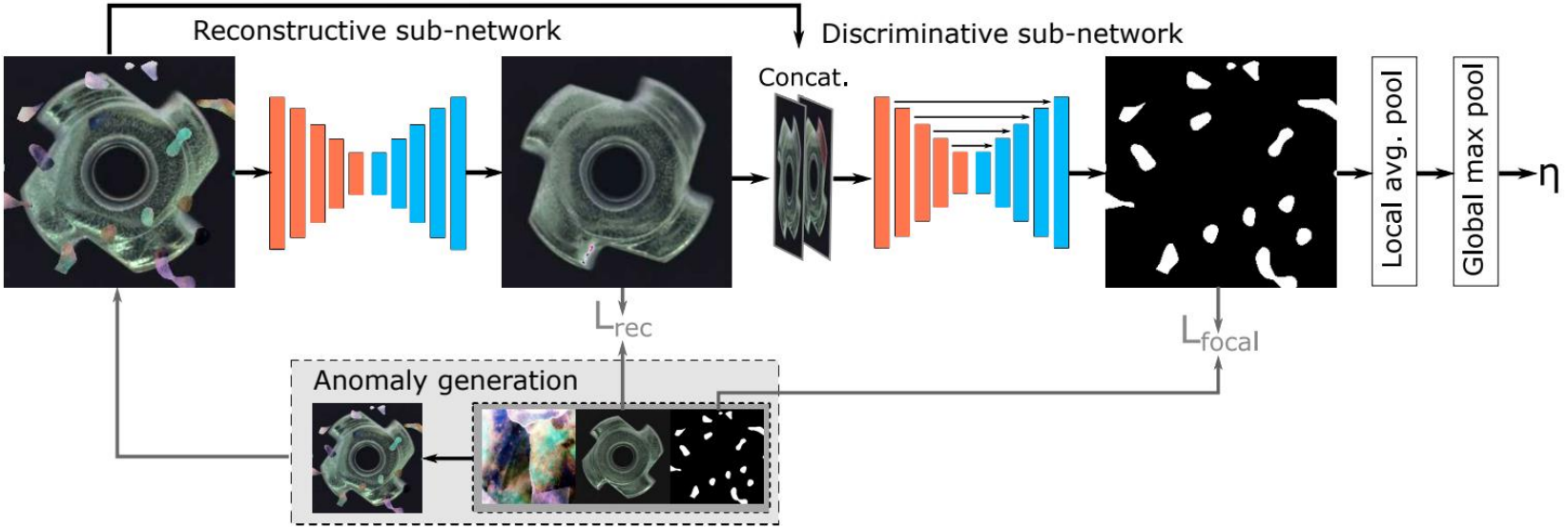
(a) Learning Self-Supervised Representation

(b) Anomaly Detection and Localization

Li, Chun-Liang, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. "Cutpaste: Self-supervised learning for anomaly detection and localization." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9664-9674. 2021.

Result 3 : Anomaly Detection in Images (2/4)

- Baselines: CutPaste [34] and **DRAEM [79]**



Zavrtanik, Vitjan, Matej Kristan, and Danijel Skočaj. "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8330-8339. 2021.

Result 3 : Anomaly Detection in Images (3/4)

- Result: The overall performance (**AUROC**) gains are close to 1%.

	Class	Localization				Detection					
		DRAEM [79]				DRAEM [79]		CutPaste [34]			
								3-way		Ensemble	
			+SSPCAB		+SSPCAB		+SSPCAB		+SSPCAB		+SSPCAB
		AUROC	AUROC	AP	AP	AUROC	AUROC	AUROC	AUROC	AUROC	AUROC
Texture	Carpet	**95.5**	95.0	53.5	**59.4**	97.0	**98.2**	**93.1**	90.7	93.9	**96.8**
	Grid	**99.7**	99.5	**65.7**	61.1	99.9	**100.0**	99.9	99.9	**100.0**	99.9
	Leather	98.6	**99.5**	75.3	**76.0**	100.0	100.0	100.0	100.0	100.0	100.0
	Tile	99.2	**99.3**	92.3	**95.0**	99.6	**100.0**	93.4	**94.0**	94.6	**95.0**
	Wood	96.4	**96.8**	77.7	77.1	99.1	**99.5**	98.6	**99.2**	99.1	99.1
Object	Bottle	**99.1**	98.8	86.5	**87.9**	**99.2**	98.4	98.3	**98.6**	98.2	**99.1**
	Cable	94.7	**96.0**	52.4	**57.2**	91.8	**96.9**	80.6	**82.9**	81.2	**83.6**
	Capsule	**94.3**	93.1	49.4	**50.2**	98.5	**99.3**	96.2	**98.1**	**98.2**	97.6
	Hazelnut	99.7	**99.8**	92.9	92.6	100.0	100.0	97.3	**98.3**	98.3	**98.4**
	Metal Nut	**99.5**	98.9	96.3	**98.1**	98.7	**100.0**	99.3	**99.6**	99.9	99.9
	Pill	**97.6**	97.5	48.5	**52.4**	98.9	**99.8**	92.4	**95.3**	94.9	**96.6**
	Screw	97.6	**99.8**	58.2	**72.0**	93.9	**97.9**	86.3	**90.8**	88.7	**90.8**
	Toothbrush	98.1	98.1	44.7	**51.0**	100.0	100.0	98.3	**98.8**	99.4	**99.6**
	Transistor	**90.9**	87.0	**50.7**	48.0	**93.1**	92.9	95.5	**96.5**	96.1	**97.3**
	Zipper	98.8	**99.0**	**81.5**	77.1	100.0	100.0	**99.4**	99.1	99.9	99.9
	Overall	**97.3**	97.2	68.4	**69.9**	98.0	**98.9**	95.2	**96.1**	96.1	**96.9**

Result 3 : Anomaly Detection in Images (4/4)

- **Anomaly localization** examples.

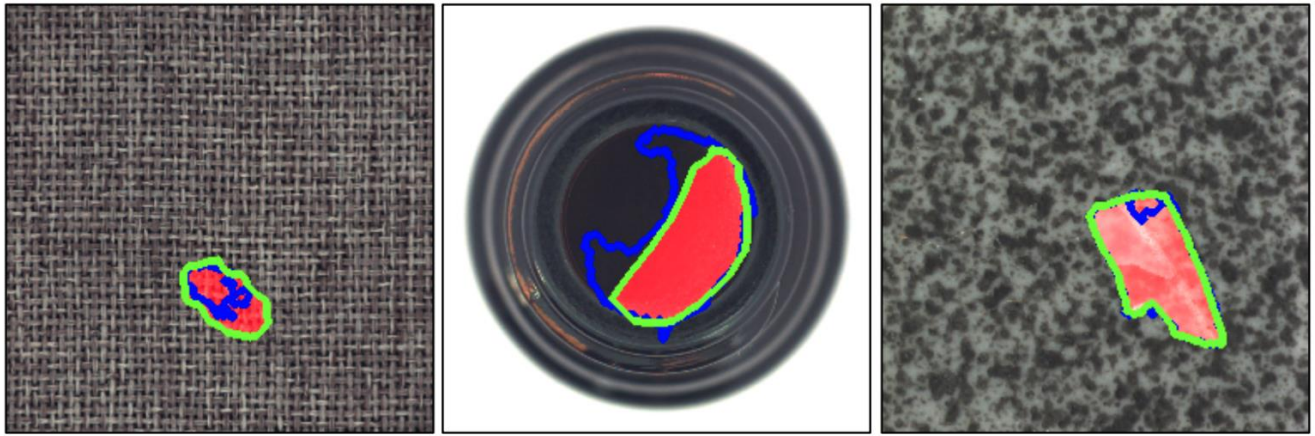
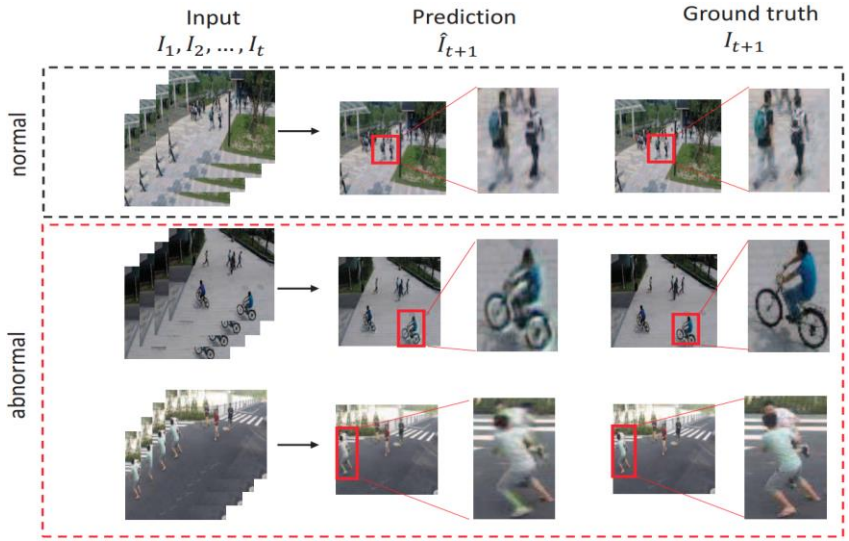


Figure 3. Anomaly localization examples of DRAEM [79] (blue) versus DRAEM+SSPCAB (green) on MVTec AD. The ground-truth anomalies are marked with a red mask. Best viewed in color.

Result 4 : Abnormal Event Detection in Video (1/2)

- Baselines: four recently introduced methods [18, **37**, 39, 49] attaining state-of-the-art performance levels in video anomaly detection.
- We integrate SSPCAB into the **auto-encoders**, not in the binary classifiers.



Result 4 : Abnormal Event Detection in Video (2/2)

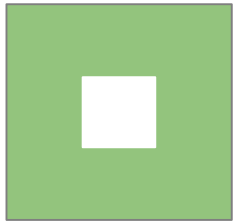
Venue	Method	Avenue				ShanghaiTech			
		AUC		RBDC	TBDC	AUC		RBDC	TBDC
		Micro	Macro			Micro	Macro		
BMVC 2018	Liu *et al.* [38]	84.4	-	-	-	-	-	-	-
CVPR 2018	Sultani *et al.* [66]	-	-	-	-	-	76.5	-	-
ICASSP 2018	Lee *et al.* [32]	87.2		-	-	76.2		-	-
WACV 2019	Ionescu *et al.* [27]	88.9	-	-	-	-	-	-	-
ICCV 2019	Nguyen *et al.* [47]	86.9	-	-	-	-	-	-	-
CVPR 2019	Ionescu *et al.* [25]	87.4	90.4	15.77	27.01	78.7	84.9	20.65	44.54
TNNLS 2019	Wu *et al.* [73]	86.6		-	-	-	-	-	-
TIP 2019	Lee *et al.* [33]	90.0		-	-	-	-	-	-
ACMMM 2020	Yu *et al.* [77]	89.6	-	-	-	74.8	-	-	-
WACV 2020	Ramachandra *et al.* [50]	72.0		35.80	80.90	-	-	-	-
WACV 2020	Ramachandra *et al.* [51]	87.2		41.20	78.60	-	-	-	-
PRL 2020	Tang *et al.* [69]	85.1		-	-	73.0		-	-
Access 2020	Dong *et al.* [12]	84.9		-	-	73.7		-	-
CVPRW 2020	Doshi *et al.* [13]	86.4		-	-	71.6		-	-
ACMMM 2020	Sun *et al.* [67]	89.6		-	-	74.7		-	-
ACMMM 2020	Wang *et al.* [72]	87.0		-	-	79.3		-	-
ICCVW 2021	Astrid *et al.* [4]	84.7	-	-	-	73.7	-	-	-
BMVC 2021	Astrid *et al.* [3]	87.1	-	-	-	75.9	-	-	-
CVPR 2021	Georgescu *et al.* [17]	91.5	92.8	57.00	58.30	82.4	90.2	42.80	83.90
CVPR 2018	Liu *et al.* [37]	85.1	81.7	19.59	56.01	72.8	80.6	17.03	54.23
CVPR 2022	Liu *et al.* [37] + SSPCAB	**87.3**	**84.5**	**20.13**	**62.30**	**74.5**	**82.9**	**18.51**	**60.22**
CVPR 2020	Park *et al.* [49]	82.8	86.8	-	-	68.3	79.7	-	-
CVPR 2022	Park *et al.* [49] + SSPCAB	**84.8**	**88.6**	-	-	**69.8**	**80.2**	-	-
ICCV 2021	Liu *et al.* [39]	89.9	**93.5**	41.05	86.18	74.2	83.2	44.41	83.86
CVPR 2022	Liu *et al.* [39] + SSPCAB	**90.9**	92.2	**62.27**	**89.28**	**75.5**	**83.7**	**45.45**	**84.50**
TPAMI 2021	Georgescu *et al.* [18]	92.3	90.4	65.05	**66.85**	82.7	89.3	**41.34**	78.79
CVPR 2022	Georgescu *et al.* [18] + SSPCAB	**92.9**	**91.9**	**65.99**	64.91	**83.6**	**89.5**	40.55	**83.46**

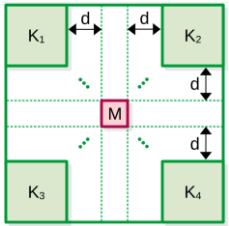
Conclusion

- SSPCAB is trained in a **self-supervised** manner, via a **reconstruction loss** of its own.
- SSPCAB is **integrated** into a series of image and video anomaly detection methods [18, 34, 37, 39, 49, 79] and obtain new **state-of-the-art** levels on Avenue and ShanghaiTech.

Discussion

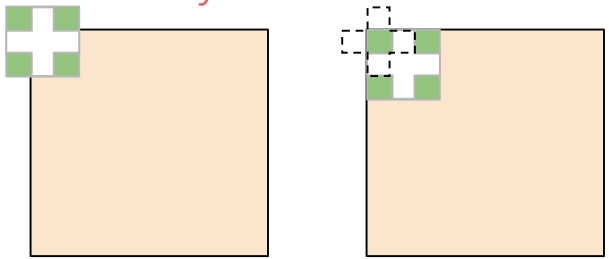
- Why receptive field look like ?





- Different from the masked reconstruction methods ?
  - **Easy to integrate**
  - Limited reconstruction ability
  - Not really mask the information

→ Put SSPCAB on the **late layer** ?



Thank You For Listening