Generalization properties of deep representations towards trustworthy AI

Elena Burceanu

Research Scientist, **Bitdefender**, Romania
University of Bucharest, Romania
Institute of Mathematics of the Romanian Academy, Romania

Trustworthy AI in Bitdefender

Computer Vision: Exploiting Space-Time Consensus in Video

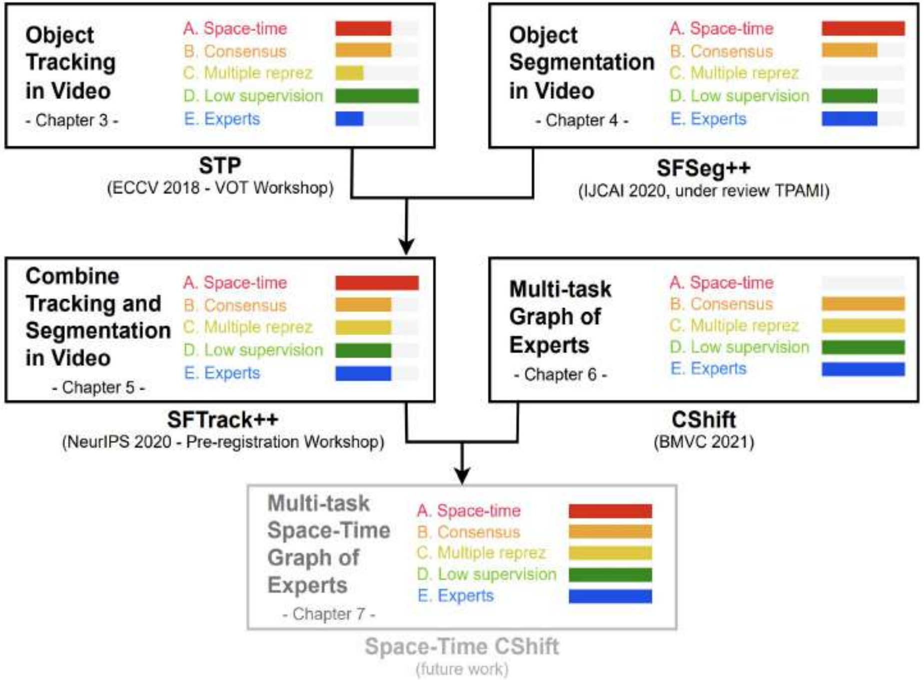
- Efficiently Exploiting Space-Time Consensus
  - Object Segmentation & Tracking in Video
  - Spectral approach

Key aspects

- Combine the spatial and temporal dimensions
- Follow consensus between complementary parts
- Learn multiple representations
- Use as many unsupervised cues as possible
- Take advantage of existing experts

=> Building more robust representations and solutions

E. Burceanu, E. Haller, M. Leordeanu



Computer Vision: DeepFake detection and localization

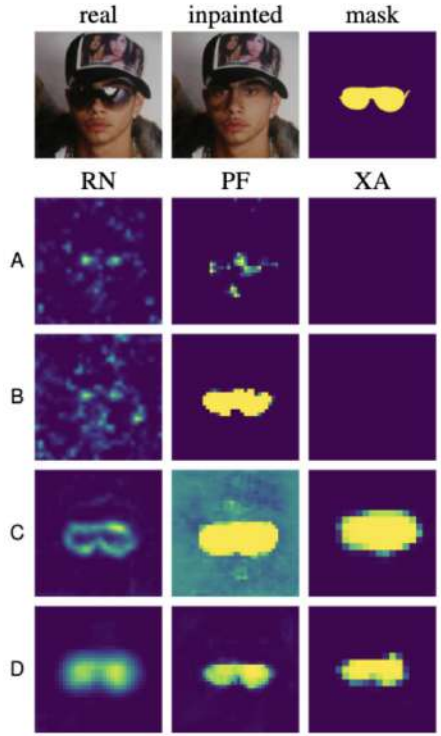
Denoising diffusion probabilistic models

- Impressive generation capabilities
- Questioning the authenticity of digital images

Detection of diffusion-generated images

- Not only a “fake” or “real” label
- But a map to indicate the manipulated area
    - Weakly-supervised

E. Oneata (Marinoiu), D. Tantaru, D. Oneata, E. Haller



NLP: Domain Adaptation for Authorship Verification

- Rethinking the Authorship Verification Experimental Setups
    - Isolate and identify biases related to the text topic and to the author's writing style
    - Explainable AI approaches guided us towards towards named entities biases
    - Models trained without them show better generalization capabilities
        - EMNLP, 2022
- VeriDark: A Large-Scale Benchmark for Authorship Verification on the Dark Web
    - Introduce a large benchmark for a new environment for Authorship Verification, DarkNet
    - Analyze the transfer learning capabilities between Authorship datasets
        - NeurIPS, Datasets and Benchmarks Track, 2022

A. Manolache, F. Brad, E. Burceanu, A. Barbalau, M. Popescu, R. T. Ionescu

Reinforcement Learning: Spectral Normalization

- RL
    - Shifts are embedded in its core definition
    - Involves interactions with an environment
    - The environment is continuously changing
    - Acquiring the ability to generalize over shifts is the key

- Spectral Normalisation for Deep Reinforcement Learning: An Optimisation Perspective
    - Regularising the value-function estimator
    - By constraining the Lipschitz constant of a layer using spectral normalisation
        - ICML 2021

F. Gogianu, T. Berariu, M. Rosca, C. Clopath, L. Busoniu, R. Pascanu

Trustworthy Anomaly Detection

through Better OOD Generalization

AnoShift - A distribution shift benchmark for unsupervised anomaly detection

Marius Drăgoi[*1] Elena Burceanu[*1,2] Emanuela Haller[*1,3] Andrei Manolache[1] Florin Brad[1]

Bitdefender, Romania[1]
bit-ml.github.io

[2]University of Bucharest, Romania
[3]University Politehnica of Bucharest, Romania

NeurIPS 2022, Datasets and Benchmarks paper

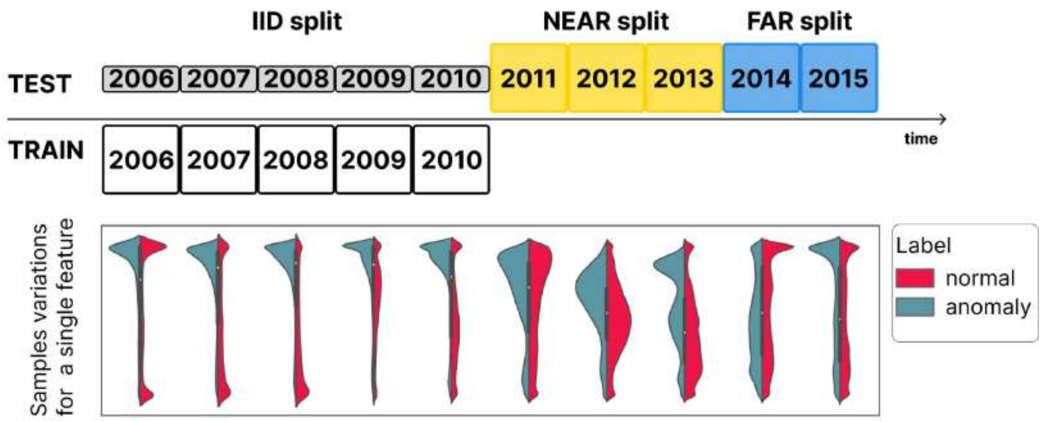
AnoShift

What we wanted

- *Continuous data* stream that spreads over a large time-span (10 years)
- The shift occurs *naturally and gradually*
- *Large* enough
- Still an *open problem* (not saturated)

Analyzed over 20 datasets: *Kyoto-2006+*

- Network traffic monitoring dataset
- Honeypots deployed in a campus
- Attacks are the anomalies



Protocol: Train on IID, test on NEAR and FAR

Key insights

We are the first to approach Anomaly Detection in distribution shift scenarios

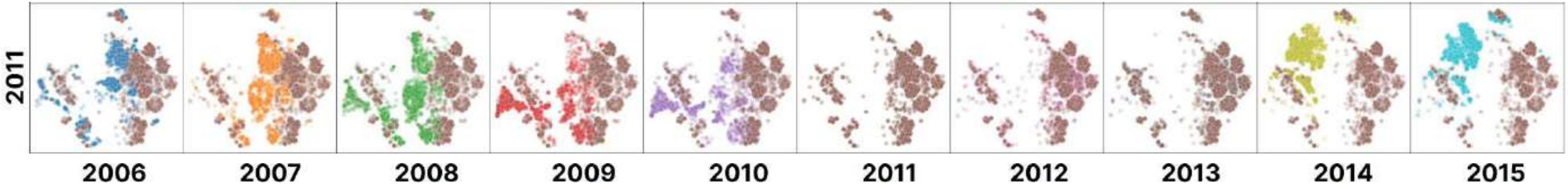
- Detailed shift analysis
  - visual representations (t-SNE)
  - per feature-level analysis
  - multi-variate distribution-level analysis (OTDD)
- AnoShift, a chronology-based benchmark
  - captures the in-time performance degradation
- Acknowledging and addressing the shift
  - to enable better anomaly detection models

Shift analysis: t-SNE

Differences in projections between years

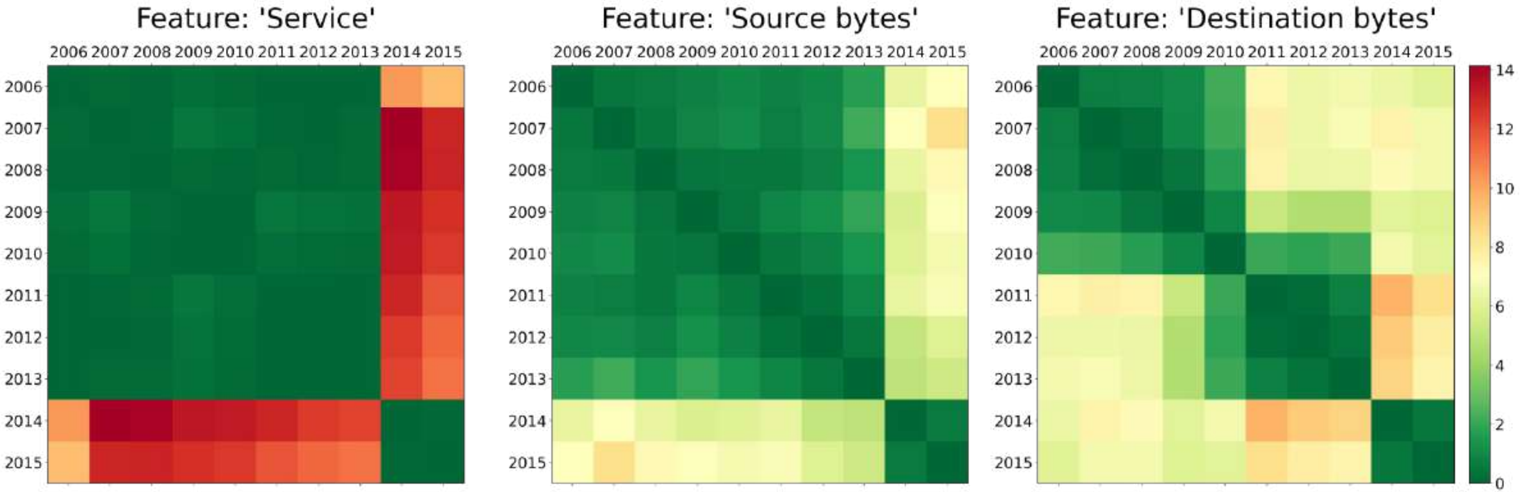
- Samples from **2011** are in **brown**
- All other years in different colors

=> Clear shifts in data distribution over the years



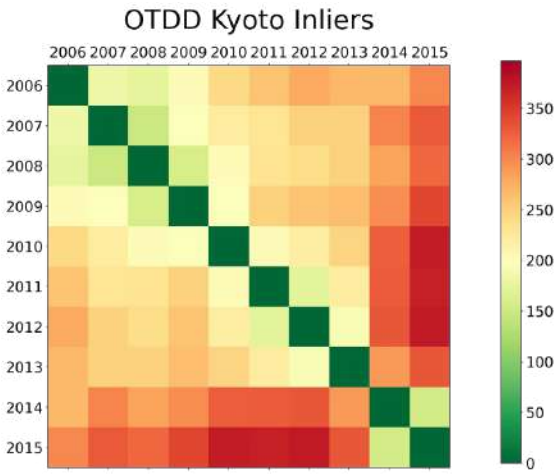
Shift analysis: feature-level

- Analyse how feature distributions change in time
- Jeffrey’s divergence between feature histograms
- Feature histogram similarity is usually higher nearby



Shift analysis: multi-variate distribution distances

- Analyse how subset distributions changes with time
- OTDD between data subsets (inliers and outliers)
- Subset distribution distance increases for inliers

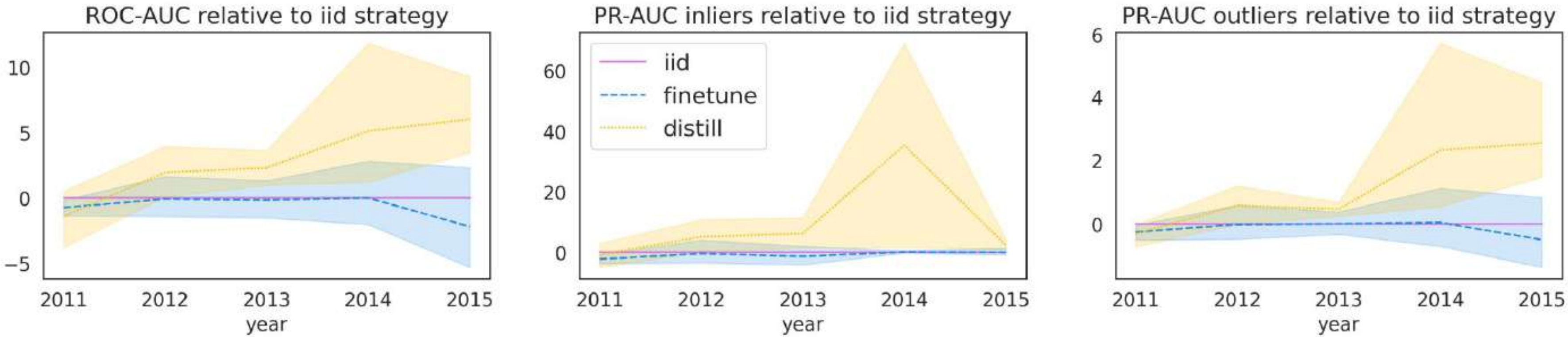


Results - ROC-AUC

- All AD models *fail to generalize* over the distribution shift
- Performance drastically drops on the FAR split

Type	Baselines	ROC-AUC ↑ IID	NEAR	FAR
Classical	**OC-SVM** [39] (train 5%)	76.86 ± 0.06	71.43 ± 0.29	49.57 ± 0.09
	IsoForest [27]	86.09 ± 0.54	75.26 ± 4.66	27.16 ± 1.69
	ECOD [24]	84.76	44.87	49.19
	COPOD [23]	85.62	54.24	**50.42**
	LOF [5]	91.50 ± 0.88	79.29 ± 3.33	34.96 ± 0.14
Deep	**SO-GAAL** [28]	50.48 ± 1.13	54.55 ± 3.92	49.35 ± 0.51
	deepSVDD [36]	$\mathbf{92.67} \pm 0.44$	$\mathbf{87.00} \pm 1.80$	34.53 ± 1.62
	AE [1] **for anomalies**	81.00 ± 0.22	44.06 ± 0.57	19.96 ± 0.21
	LUNAR [14] (train 5%)	85.75 ± 1.95	49.03 ± 2.57	28.19 ± 0.90
	InternalContrastiveLearning [41]	84.86 ± 2.14	52.26 ± 1.18	22.45 ± 0.52
	BERT [11] **for anomalies**	84.54 ± 0.07	86.05 ± 0.25	28.15 ± 0.06

Addressing the distribution shift



Training strategies

1. iid: a new model for each interval
2. finetune: finetune over previous year
3. distil: distillation from the previous year

Insights

- Distillation performs the best (+3%)
- Better modelling of inliers (higher PR-AUC for inliers)

Env-Aware Anomaly Detection: Ignore Style Changes, Stay True to Content!




Ștefan Smeu[*1,2] Elena Burceanu[*1] Andrei Nicolicioiu[3] Emanuela Haller[1]

Bitdefender, Romania[1]
bit-ml.github.io

[2]University of Bucharest, Romania
[3]MPI for Intelligent Systems, Tübingen

NeurIPS 2022, Workshop on Distribution Shifts paper

Key insights

Same focus: Unsupervised Anomaly Detection in non-stationary distributions

- Benchmark for images
  - As opposed to tabular data like in AnoShift
- Split the data in environments: Env-aware learning methods in pretraining
  - Produce better embeddings for Anomaly Detection
- EA-MoCo method
  - Adjusting contrastive learning to be aware of multiple environments improves the performance even over supervised approaches

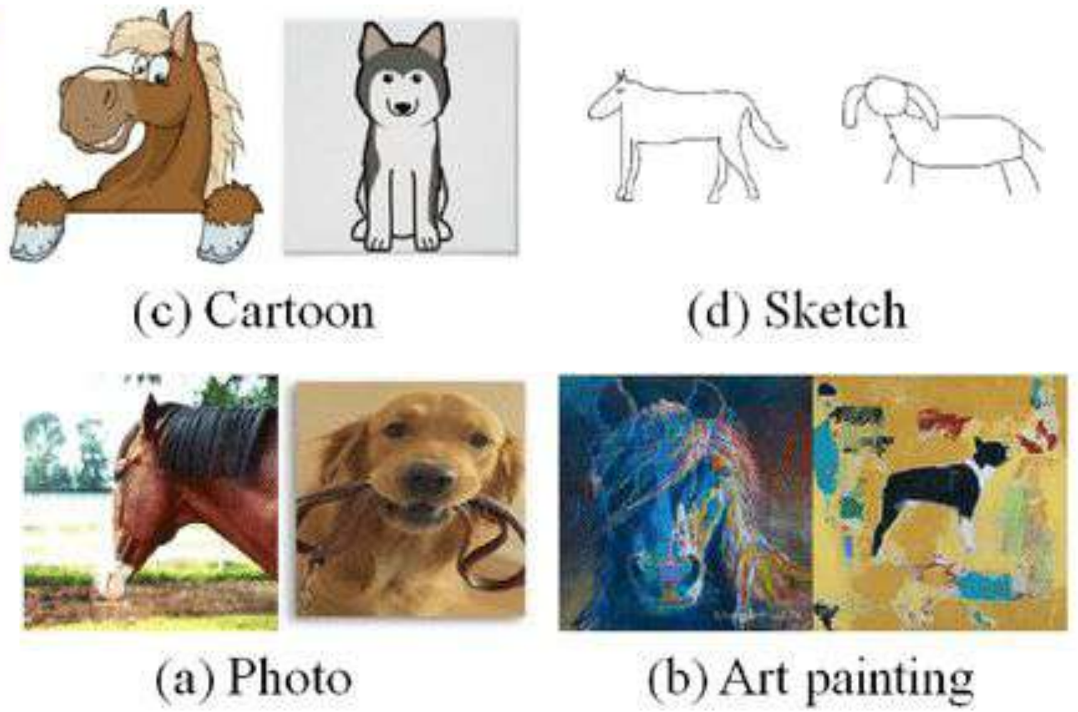
Robust to Style changes, but detect Content changes as Anomaly

Style environments:

- cartoon, sketch, photo, art painting

Content classes:

- horse and dog



Out-of-distribution regimes (test time)

- 4 different scenarios for train vs test distribution changes
- Differentiate between
  - Style vs
  - Content changes

Our scenario

- Style is OOD
  - we want to ignore this
  - to be robust to it
- Content is OOD => detect as Anomaly

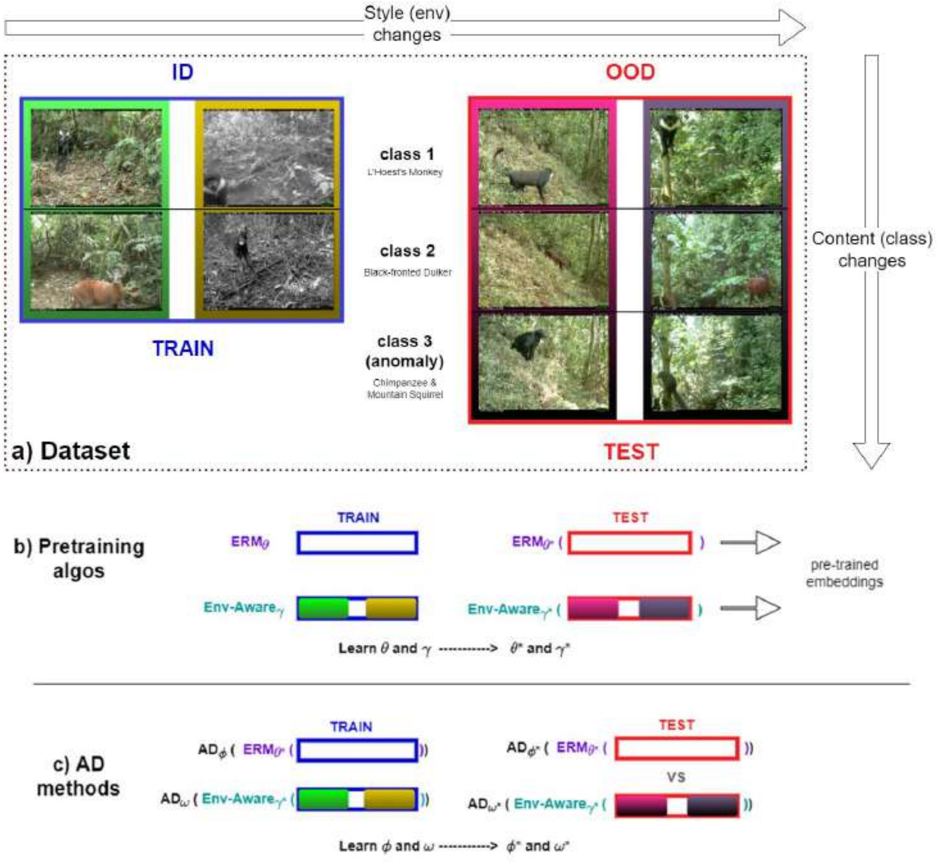
	Style	Content	Description
A.	ID	ID	**Assumption:** $p_e(x_S, x_C, y)$, $p_e(x_S, x_C)$ are constant
			Goal/Task: model $p_e(y\|x)$ or $p_e(x, y)$ or $p_e(x)$
			algorithms following the ERM paradigm
B.	OOD	ID	**Assumption:** $p_e(x_S)$ changes over envs - closer to real-world scenarios
			Goal/Task: same as **A.**, while being robust to Style changes
			IRM, V-Rex, Fish, Lisa
C.	ID	OOD	**Assumption:** $p_e(x_C)$ changes over envs
			Goal/Task: detect Content changes
			open set recognition; detect semantic anomalies or novelties
D.	OOD	OOD	**Assumption:** both $p_e(x_S)$, $p_e(x_C)$ change over envs - closer to real-world scenarios
			Goal/Task: same as **C.**,while being robust to Style changes
			EA-MoCo (our approach)

Anomaly Detection Setup

Learning process

1. Learn embeddings robust to style changes
   a. Supervised, using env-aware methods
   b. **Unsupervised**, **EA-MoCo**, an env-aware contrastive approach

2. Anomaly detection using those learned embeddings

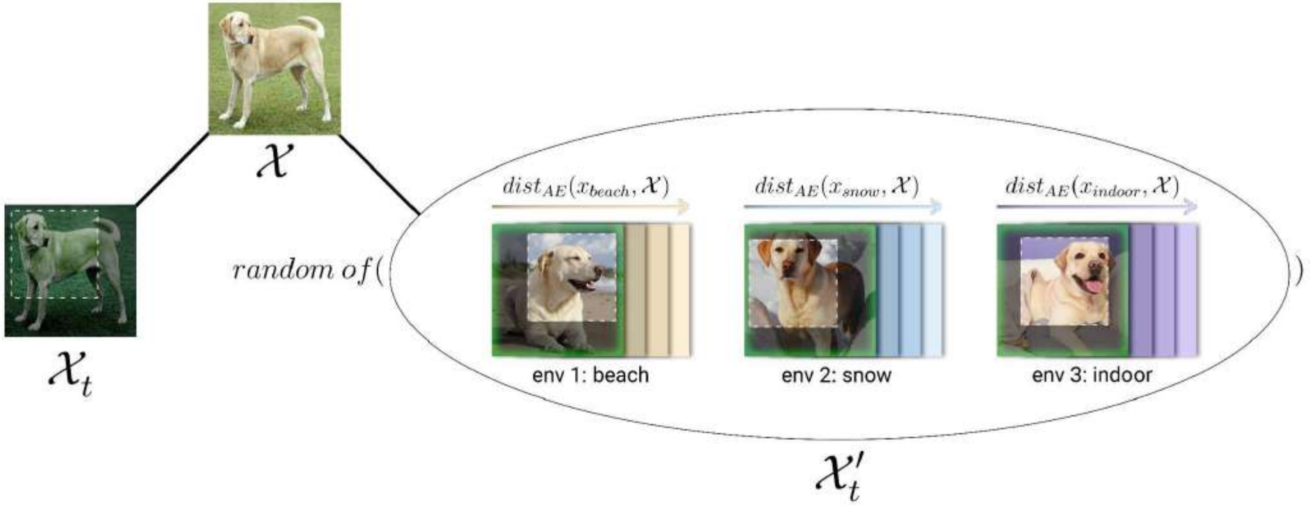


EA-MoCo - strategy for positive pair selection

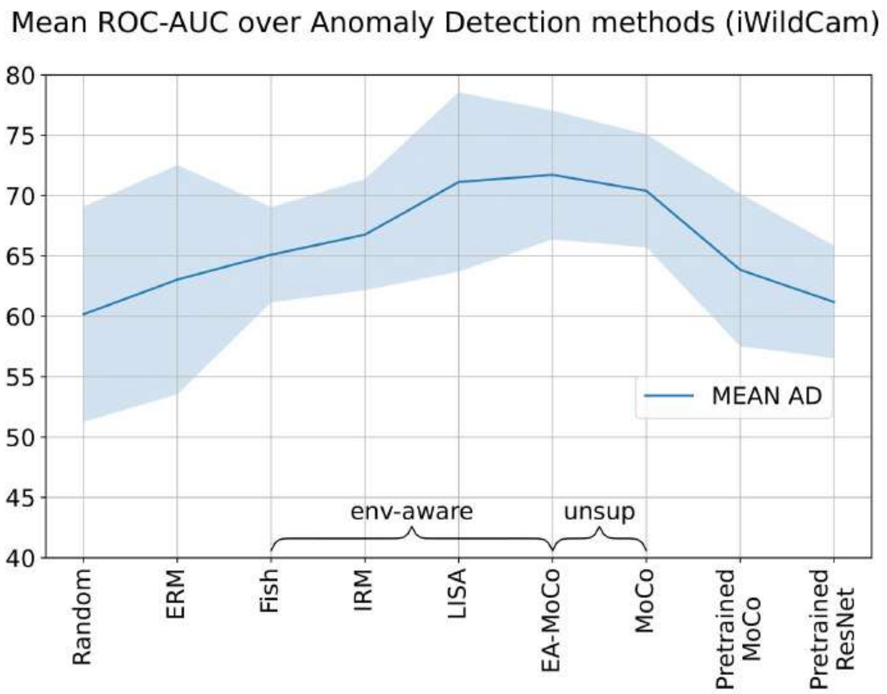
Positive pair is formed of:

- usual, random augmented version of anchor ($\mathcal{X}_t$
- closest sample from a different, random environment w.r.t. a trained autoencoder embeddings ($\mathcal{X}'_t$

Takeaway: Style (environment)-aware pretraining when building the positive samples!



Results



Pretrain	None	Supervised				Unsupervised		Other dataset	
	Random	ERM	Fish	IRM	Lisa	**EA-MoCo**	MoCo v3	MoCo v3	ResNet
IsoForest	65.2	63.1	68.0	64.3	**75.2**	70.9	68.4	64.6	61.8
INNE	50.1	67.7	66.1	68.7	76.5	**77.0**	71.9	68.7	57.8
LODA	65.1	63.8	66.7	66.2	**73.9**	71.1	66.9	67.1	69.9
OCSVM	57.9	67.5	65.5	64.5	**78.4**	71.4	68.5	57.1	62.1
PCA	64.1	40.4	63.3	64.4	55.6	**67.7**	63.9	60.9	63.2
LOF5	43.2	61.0	59.7	61.3	65.1	60.9	**68.3**	58.5	53.2
KNN	73.2	75.7	72.0	77.7	66.9	77.0	**78.9**	76.5	57.8
KDE	62.6	65.1	59.4	67.0	77.4	**77.8**	76.3	57.4	63.6
Mean AD (OOD)	60.2	63.0	65.1	66.8	71.1	**71.7**	70.4	63.8	61.2

Anom. Detect. method

- Env-aware methods perform better
- EA-MoCo scores best on most AD methods
    - And it is fully unsup!

Takeaway message

- Distribution shift of the data
  - A serious problem for ML models (and for “trusting” AI)
  - We are the first to address it in the unsupervised scenario, for Anomaly Detection
- AnoShift benchmark
  - Tabular data, network traffic
  - Large data, spans over 10 years, continuous data that gradually changes over time
- Env-Aware MoCo
  - Define anomalies from the content vs style point of view
  - Env-aware pretrainig helps
  - Propose an env-aware unsupervised pretrainig

Thank you! Questions?

eburceanu@bitdefender.com

bit-ml.github.io

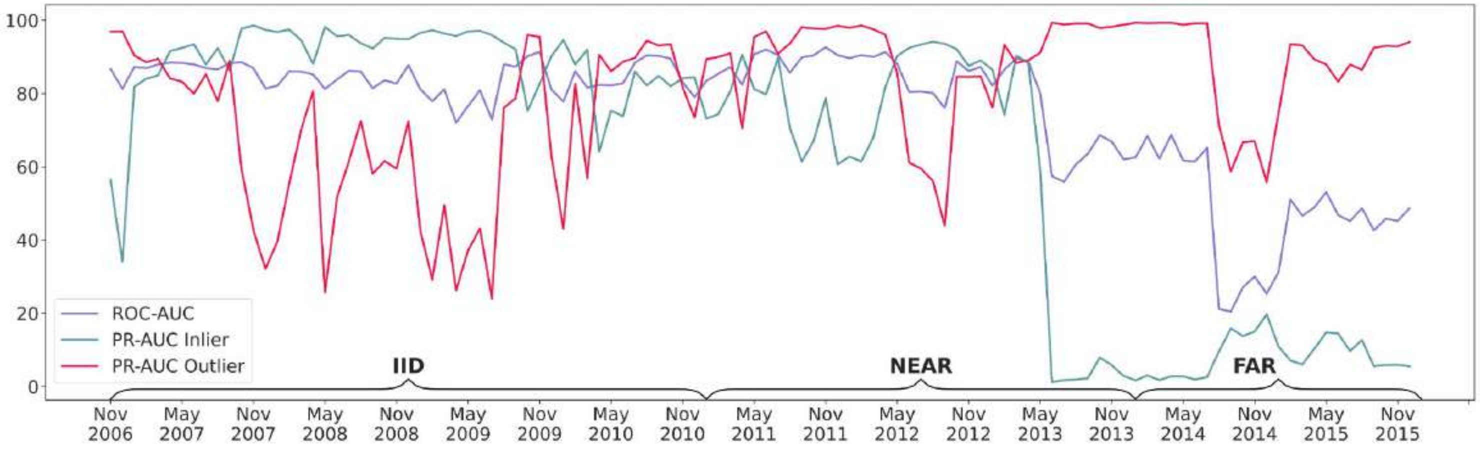
BERT for anomalies

- Train in MLM mode
- Anomaly score based on masked token retrieval probabilities

$$anomaly\_score([w_1, w_2, ..., w_t]) = \frac{\sum_{i=1..n} \sum_{j=1..t}^{mask_i \sim Masks_t^p} (1 - P(\hat{w}_j{}^i))}{n}$$

$$P(\hat{w}_j{}^i) = \begin{cases} 1, & \text{if } mask_i(\text{j}) = 0 \\ P_M(w_j | \theta_M, [\hat{w}_1{}^i, ..., \hat{w}_t{}^i]), & \text{if } mask_i(\text{j}) = 1 \end{cases}$$

Results



Monthly performance

- Modeling the **inliers**: IID > NEAR > FAR
- Poor modeling for the **outliers**