Applications of Representation Learning

Retrieval

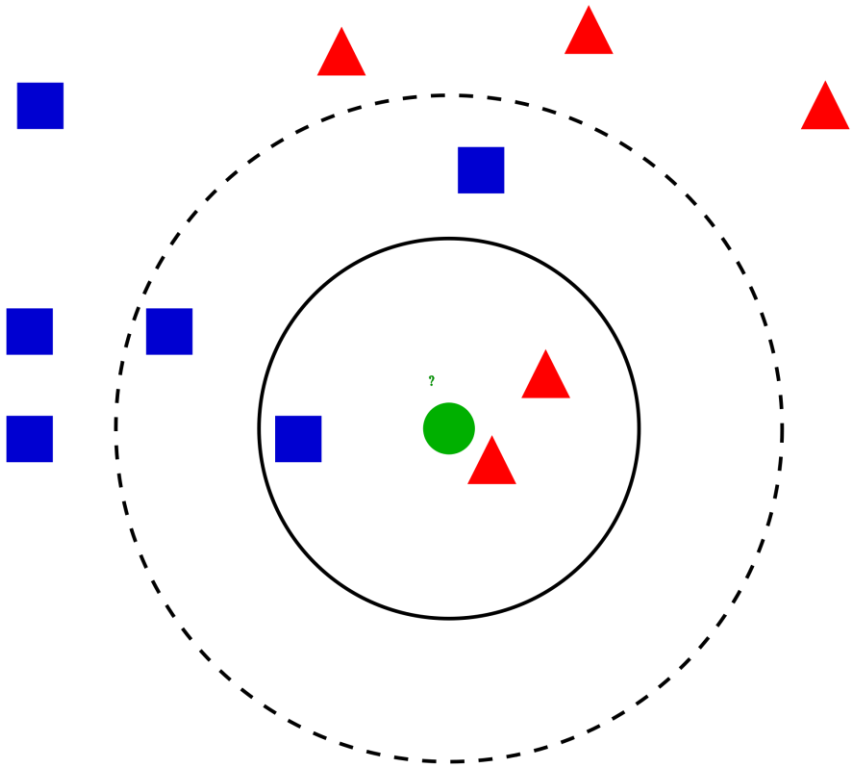
- One of the most immediate applications is retrieval
- Given image x, find the K most similar images in dataset D
- Poorly defined, similar according to which attribute?

Semantic Representation

- We have already tackled this issue, two options:
  - Use supervision on labelled similar images
  - Use generative augmentation to tell us which attributes don’t count
- Method from previous weeks learn image to representation encoder

Retrieval in Practice

- Simple:
  - Use encoder to extract representations for all images
  - Compute K nearest neighbors
- Choice of metric matters (L1, L2, cosine)
- Hubness problem – some representations similar to most



Anomaly Detection

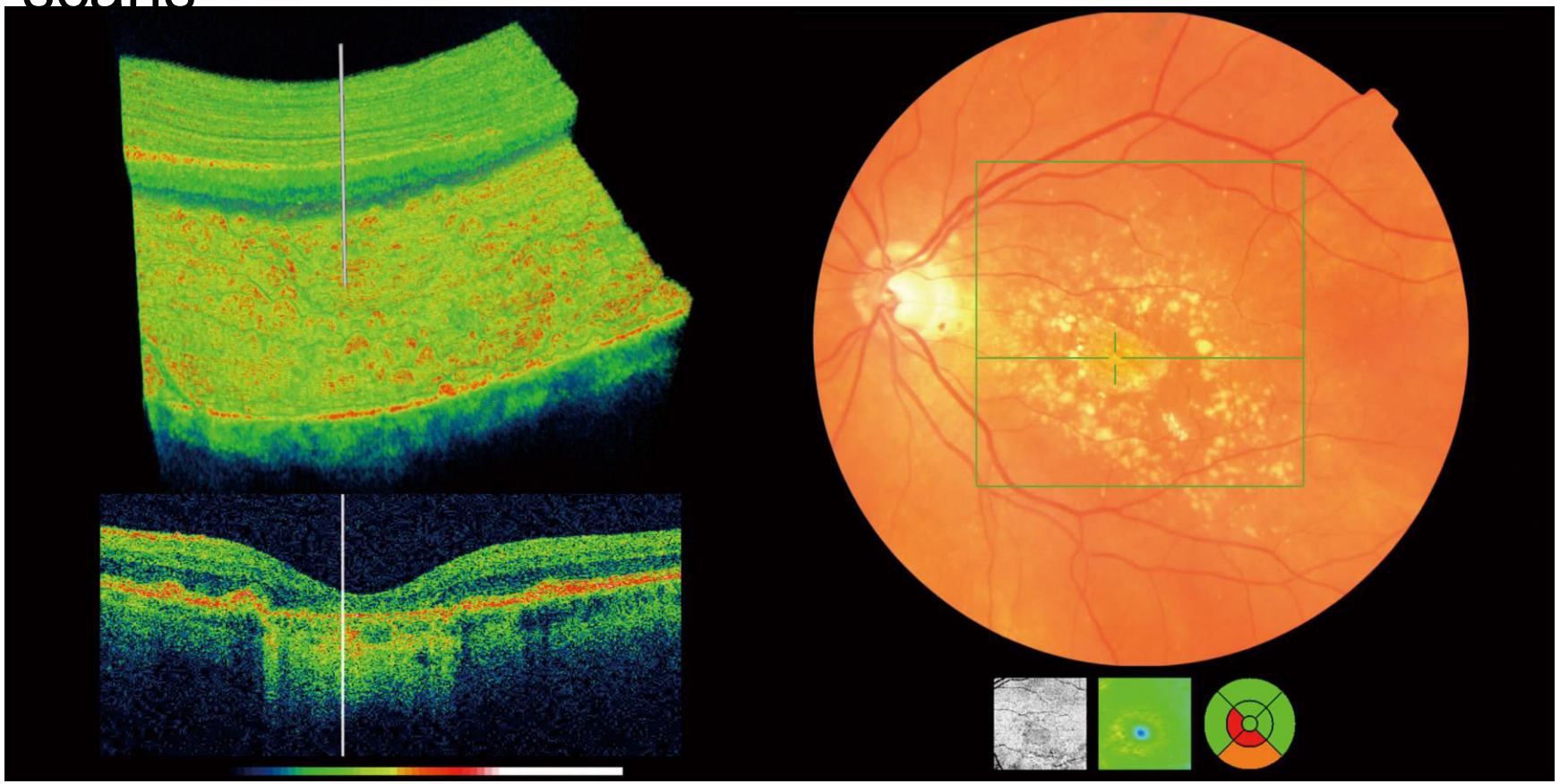
Objective: Discovering *unique*, *interesting* phenomena

Why is it important? – Cyber and Manufacturing

- Cyber security: unusual program signatures, network activity
- Production line: discovering faulty products

Why is it important? – Medicine

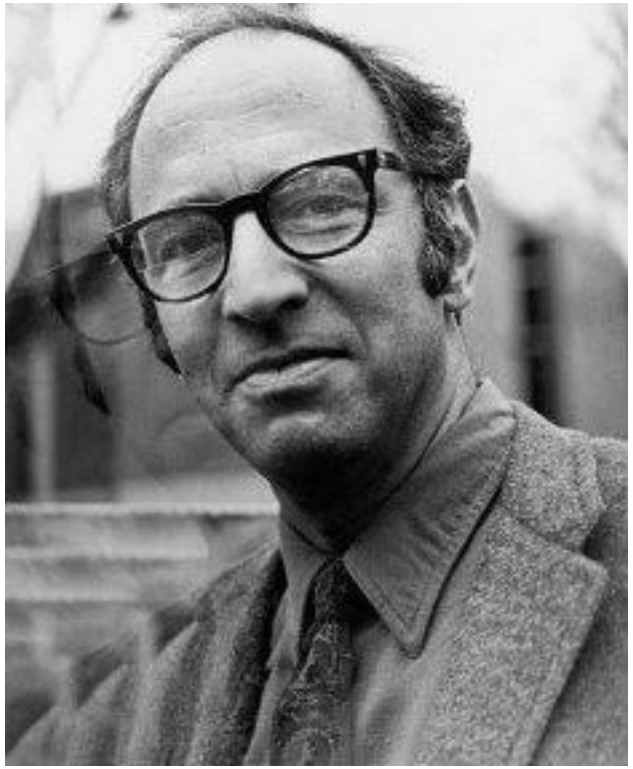
- Finding biomarkers for early discovery of disease in routine medical scans



Anomaly detection is key for discovery

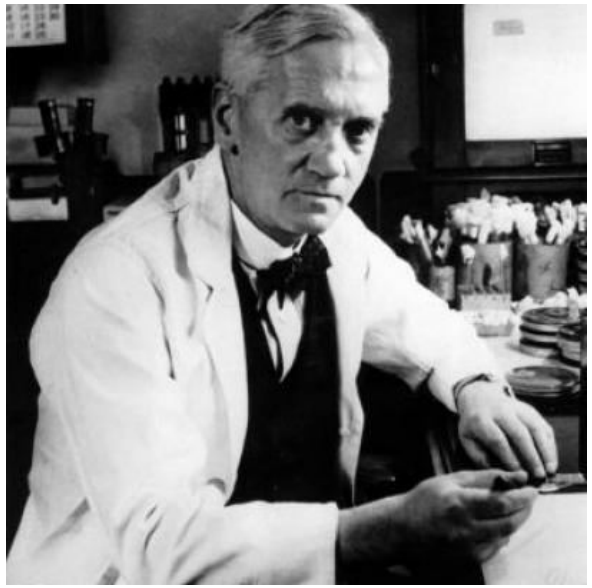
Thomas Kuhn, The Structure of Scientific Revolutions (1962)

- *“Discovery begins with the awareness of anomaly”*
- *“The paradigm change is complete when the paradigm has been adjusted so that the anomalous become the expected. “*
- *The result is that the scientist is able to see nature in a different way”*



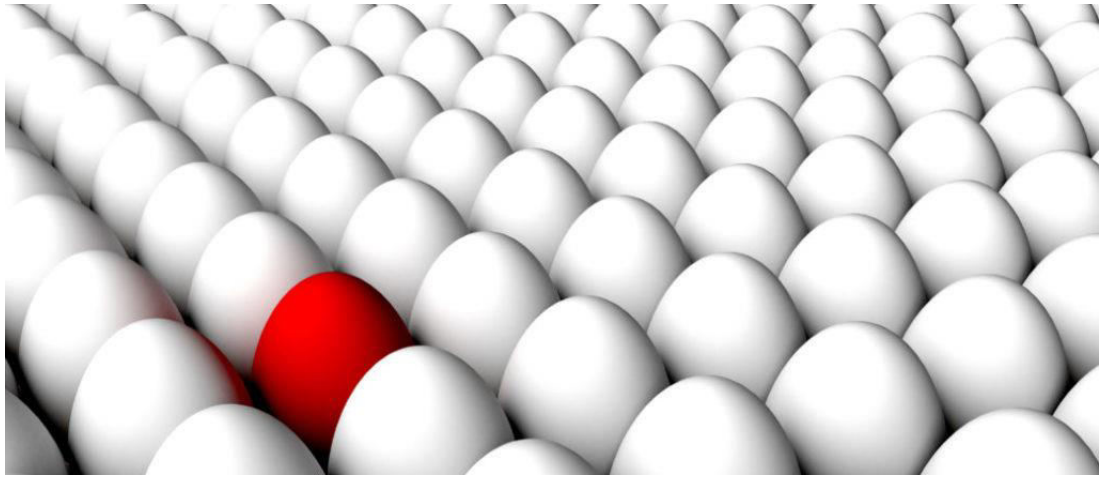
Discovery of Penicillin

- Fleming returned from vacation and inspected his petri dishes
- Dishes contained bacteria everywhere except for for moldy areas
- This anomaly lead to the discovery of Penicillin



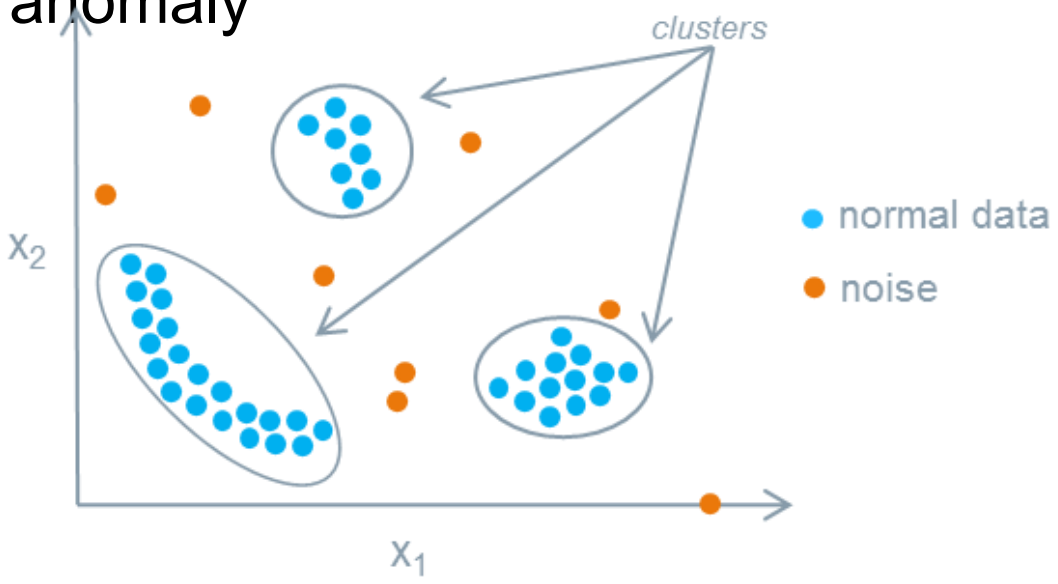
Standard Setting

- Task: Detect if a new sample is different from previously seen
- Training stage: Observe normal samples and train model
- Inference: Observe new sample and decide:
    - Similar to previously seen? - Normal
    - New pattern? - Anomaly



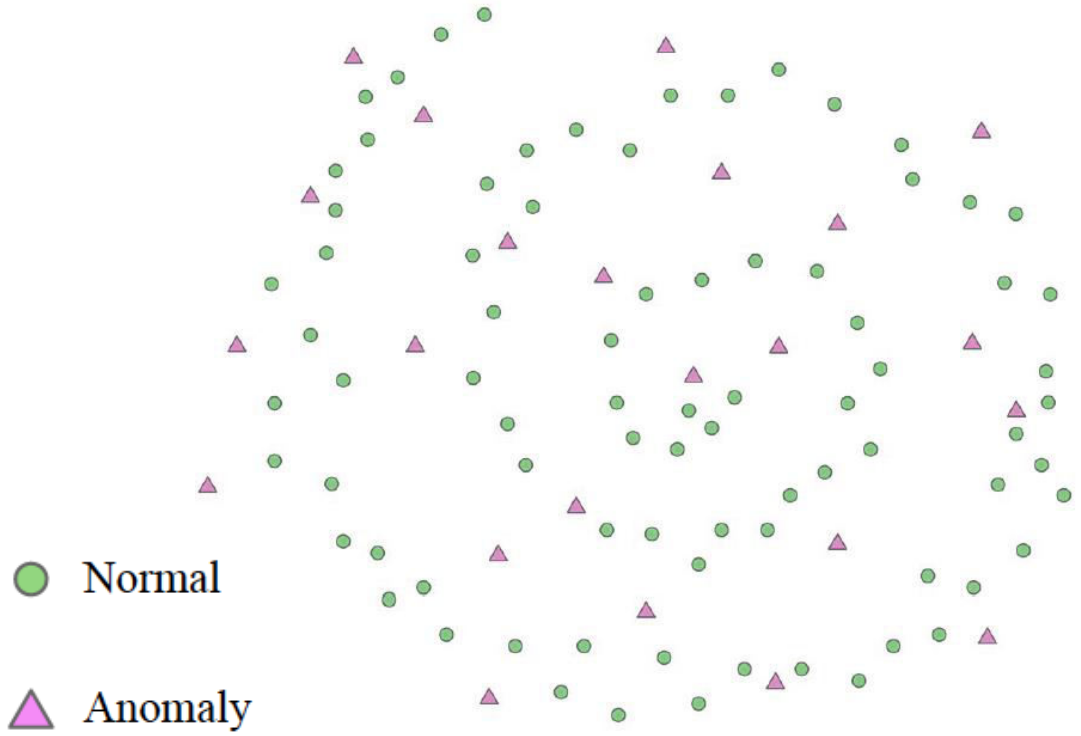
Reformulating Anomaly Detection as Density Estimation

- One definition of anomaly: sample that lies in a low-density region of space
  - Intuition: few samples are similar to the anomaly
- Classifier:
  - Low probability (p(x) < th): Anomaly
  - High probability (p(x) > th): Normal



Challenge #1: Estimation is Hard

- Estimation of probabilistic models from a small sample size is hard
- A key objective of statistics
- Particularly hard for high-dimensional, non compact data.
- We have learned many methods…

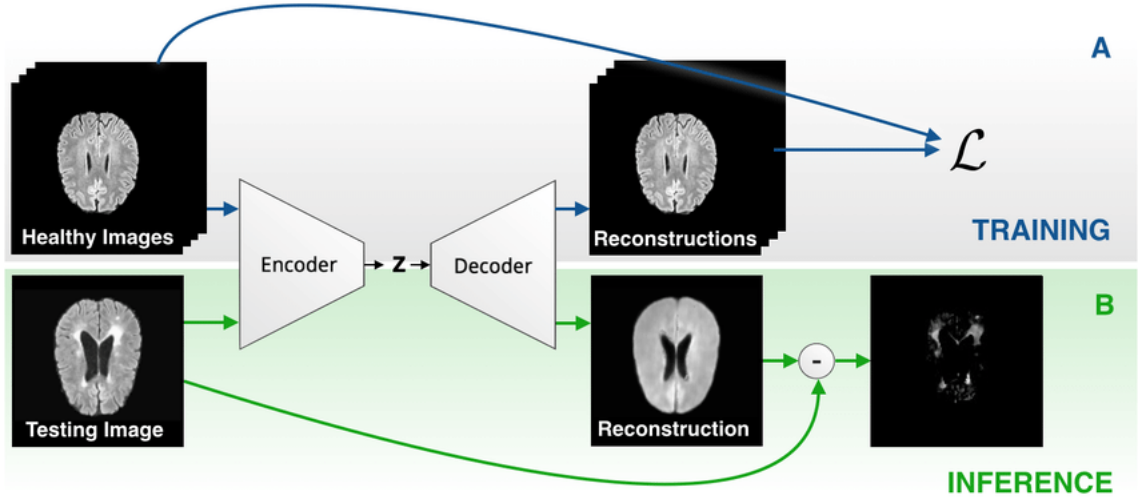


Challenge #2: Representation Must Capture Interesting Variation

- Consider a representation that contains two parts (c, u)
- c is correlated with semantic anomalies, u is correlated with nuisance attributes
- E.g. representations of production line images may have two parts:
  - c: presence of cracks and scratches
  - u: different orientation of camera, we consider this attribute as nuisance
- An unusual value of u will trigger an anomaly, but is not semantically interesting

Out-of-Distribution Generalization

- Out-of-distribution generalization (OOD)
- Idea:
  - train a classifier/autoencoder on normal data
  - At test time evaluate accuracy of model
  - Assumption: model generalizes better on normal (in-distribution) data than on OOD

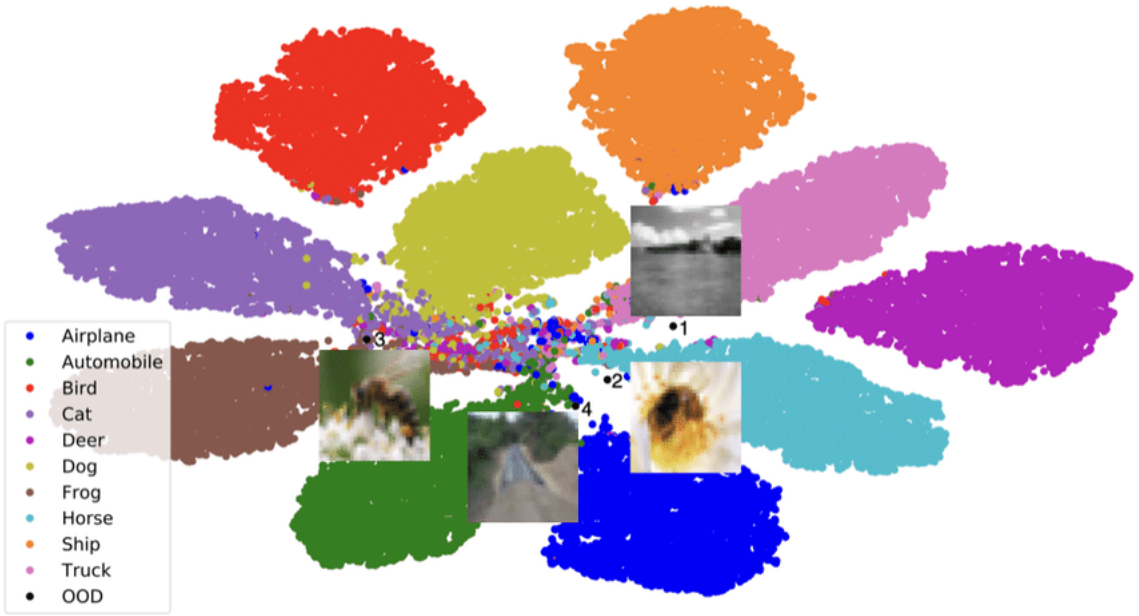


Most Deep AD Focus on Poor OOD Generalization

- Deep learning methods famously generalize poorly out-of-distribution
- Best case: classifier performs well in-distribution, poorly OOD
- Limitation: deep networks don't always generalize poorly OOD

Our Approach: Representation is All You Need

- Effective deep representation can overcome both classical limitations:
  - Density estimation is easier due to reduced dimension and denser support
  - More informative representation – more correlated to semantic anomalies

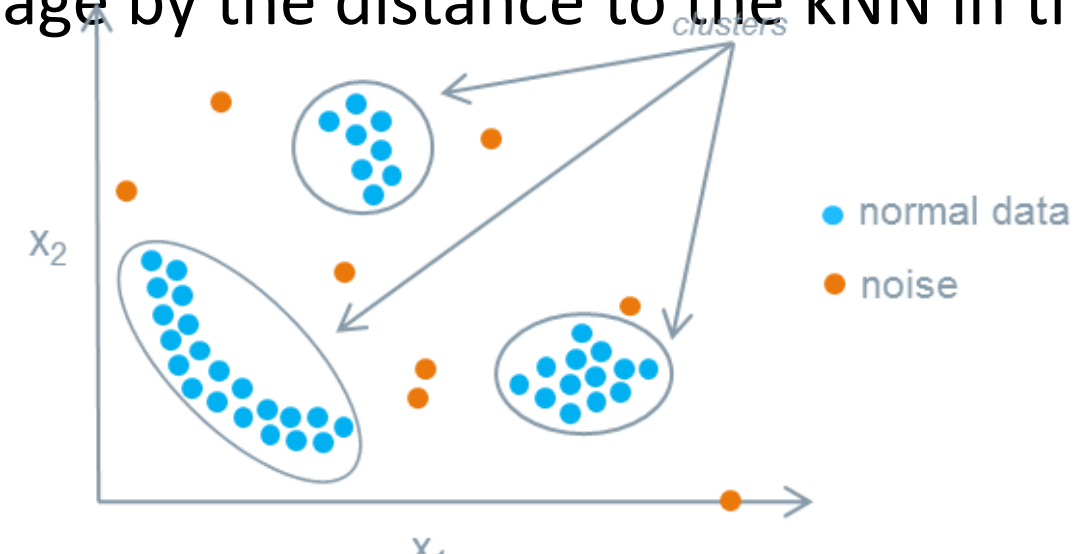


Which Representation Encoder?

- Previous: self-supervised representations trained only on normal data
  - E.g. autoencoders, rotnet, contrastive learning
- Missed opportunity - huge external resources exist e.g. ImageNet dataset
- Our approach: use external (unrelated) data to pretrain power representation encoders

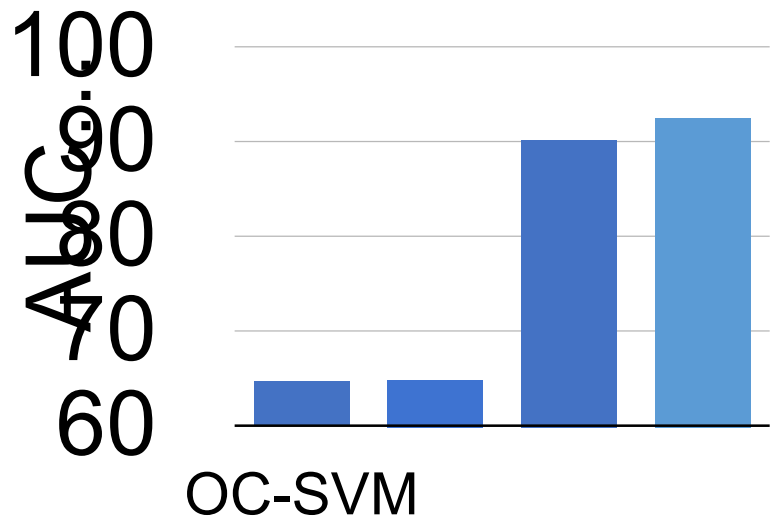
Experiment #1: Deep Nearest Neighbors

- How powerful are deep representations coupled with trivial density estimation?
- Extracted deep representation from each image using a ResNet trained on Imagenet
- Scored each test image by the distance to the kNN in the training set in feature space



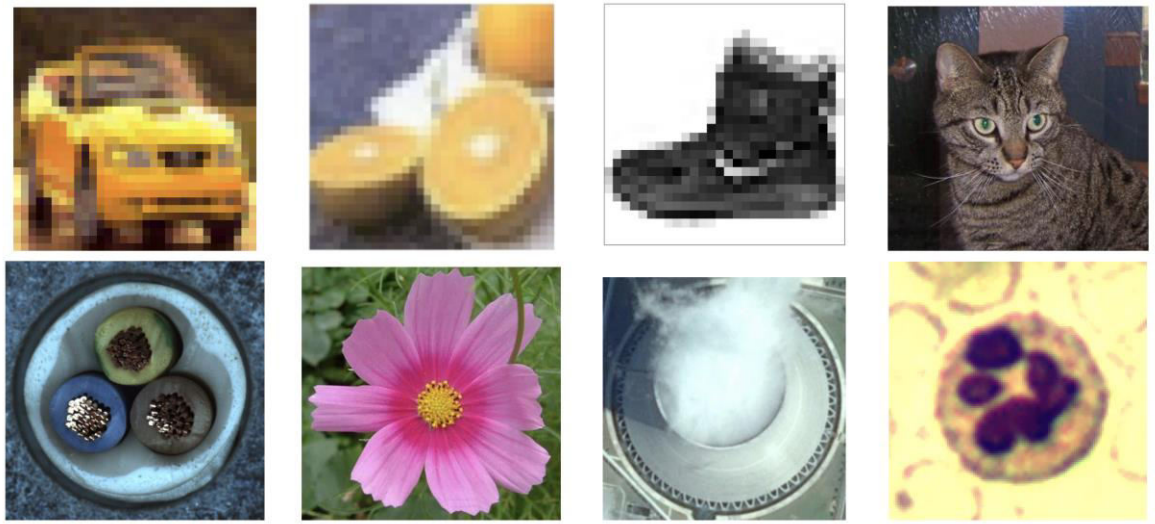
Experiment #1: Results

- DN2 achieves better results that much more complex SOTA methods
- Representation quality is more important than the statistical model



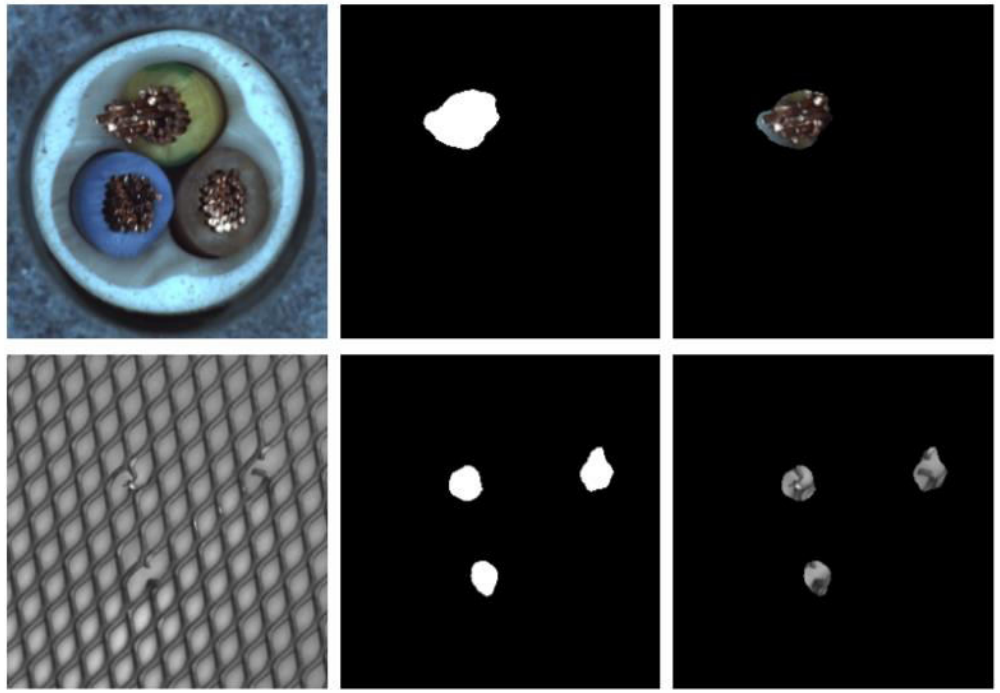
Experiment #1: Generality Across Datasets

- ImageNet-pretrained feature encoders generalize well across many image datasets
- We could even use it to segments foreign language into words (Fuchs et al., Interspeech'22)



Experiment #2: Deep Anomaly Segmentation

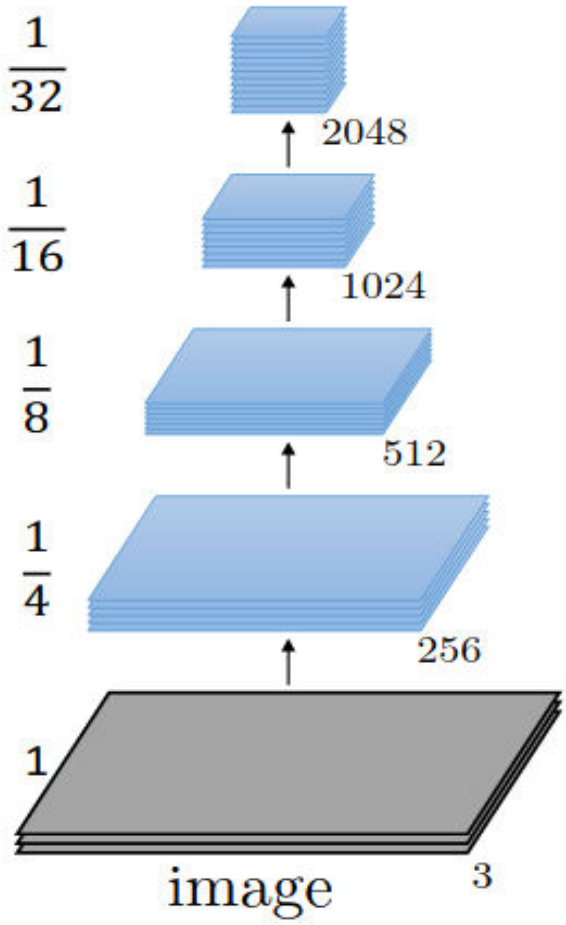
- Task: label all image pixels as normal or anomalous



Cohen, and Hoshen. Sub-image anomaly detection with deep pyramid, 2020

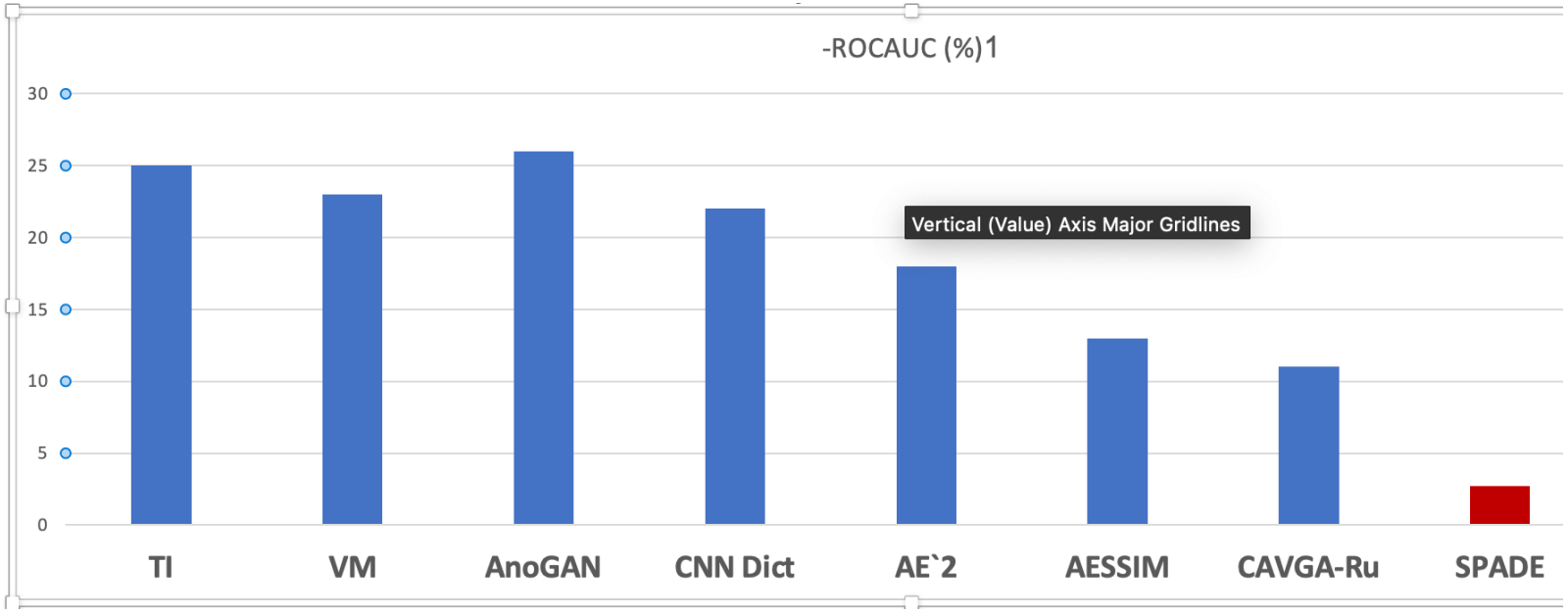
Experiment #2: Deep Anomaly Segmentation

- Method: same as DN2, but a representation is extracted for every pixel
- We named the method SPADE
- Per-pixel features extracted by concatenating different later activations
- Encoder is pretrained on ImageNet
- kNN is performed vs. all the pixel representations in the train set



Experiment #2: SPADE Results

- SPADE performed much better than previous work
- True both for image and pixel level on the MVTec datasets
- Followup method PatchCore virtually solves the dataset



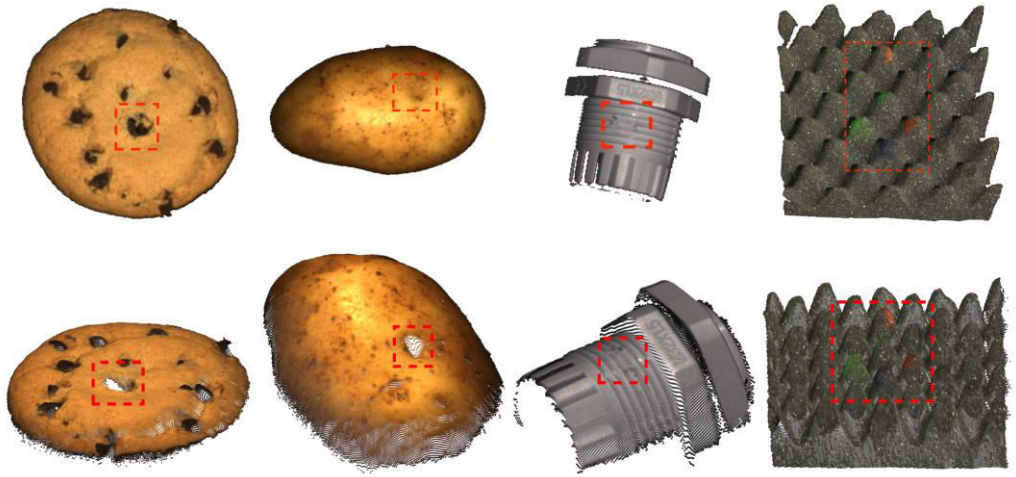
The Best Representation is Not Always Deep

- We found that simple handcrafted representations are sometimes SoTA:
  - Time series: random projection features of each window
  - 3D point clouds: classical rotation-preserving features (e.g. FPFH)
  - Tabular data: raw data (deep or handcrafted features did not improve – when no prior)

An Empirical Investigation of 3D Anomaly Detection and Segmentation, Horwitz and Hoshen, 2022

Time Series Anomaly Detection by Cumulative Radon Features, Tzachor and Hoshen, 2022

Perspective Paper, Anomaly Detection Requires Better Representations, Reiss, Cohen, Horwitz, Abutbul, Hoshen, ECCVW 2022



Story So Far

- Anomaly detection requires a representation such that:
  - The PDF of normal data is easy to estimate
  - Small overlap between PDFs of normal and unknown anomaly distributions
- This *might* be impossible in the adversarial case
- In some important settings (presented here), a good representation may be *guessed*

Desiderata for Anomaly Detection Methods

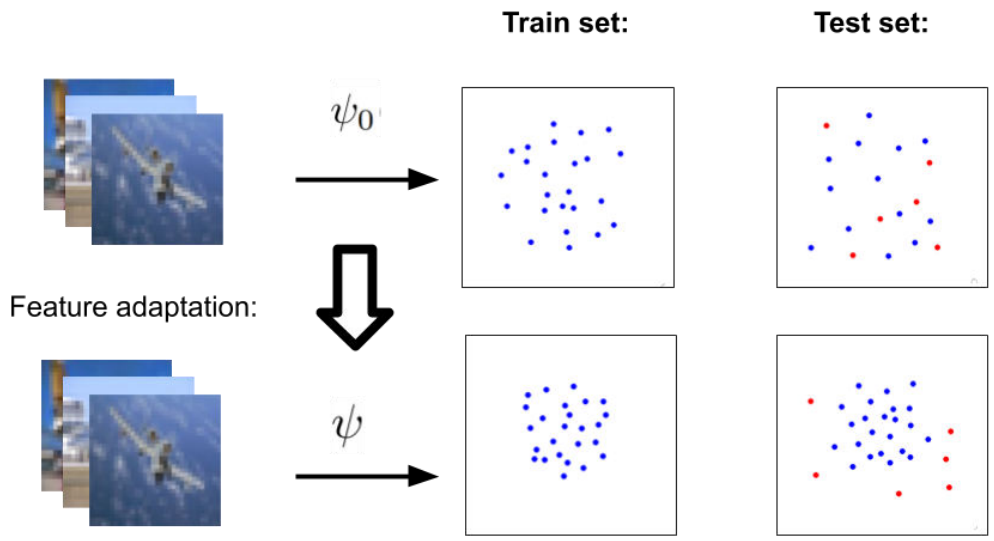
1. Should be able to use all available data i.e., external and normal training data
2. Should be interpretable
3. Should be able to use guidance to resolve ambiguity
4. Should extend to all data modalities

Desideratum 1: Combining External and Internal Data

- Anomalies are unexpected and not known at training time, but normal data are!
- Previous SoTA: self-supervised training on normal training set only
- Our experiments: externally pretrained encoders
- Our challenge: can we combine pretraining with self-supervised adaptation?

PANDA - Pre-trained ANomaly Detection Adaptation

1. Pretrain representation encoder on an external large-scale dataset (ImageNet)
2. Fine-tune pre-trained ResNet features using the compactness criterion
3. Use kNN scoring on the representations as before

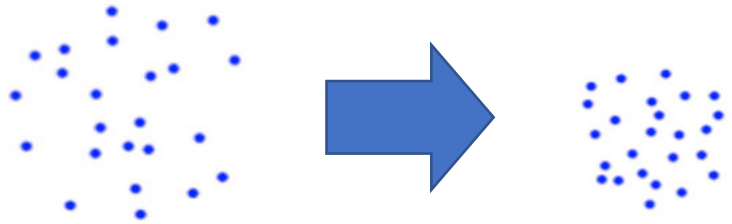


Reiss, Cohen, Bergman, Hoshen, Panda: Adapting pretrained features for anomaly detection and segmentation, *CVPR 2021*

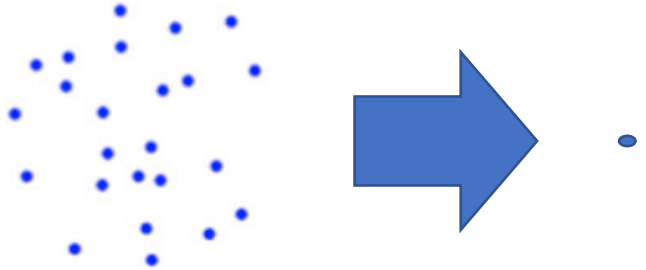
Adapting Pre-trained Features

- Compactness criterion for fine-tuning features

$$L_{compact} = \sum_{x \in \mathcal{D}_{train}} \|\psi(x) - c\|^2$$

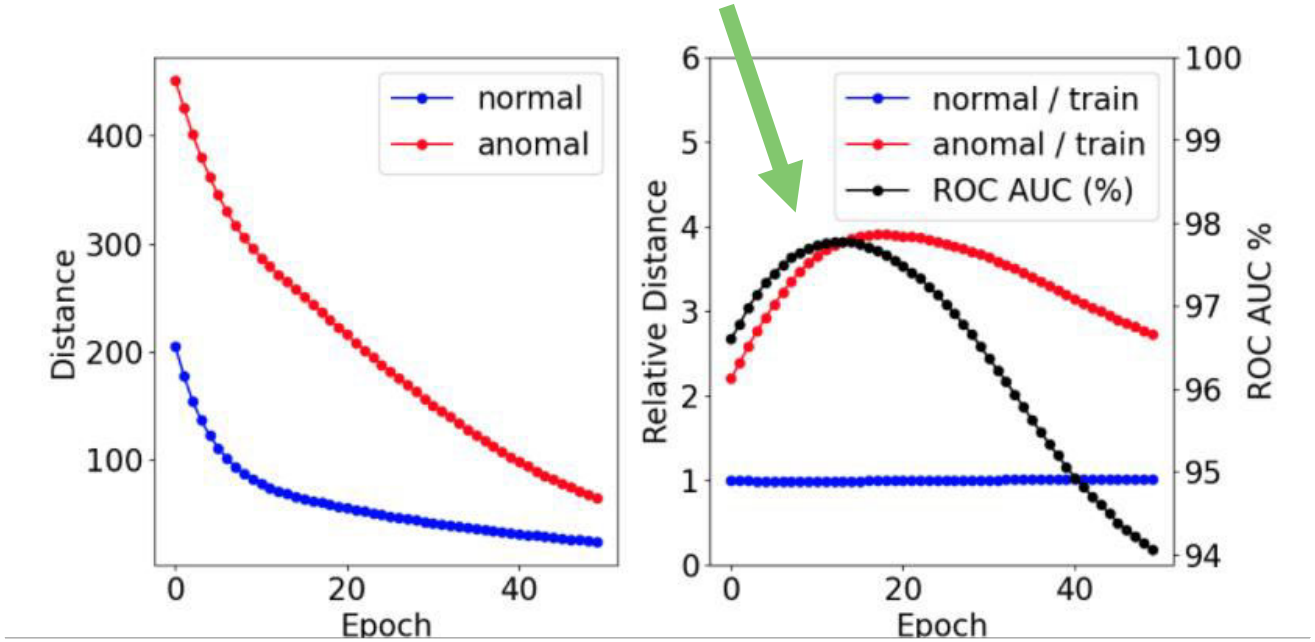


- Will this not collapse to a singular point? Catastrophic collapse?



What about the feature collapse problem?

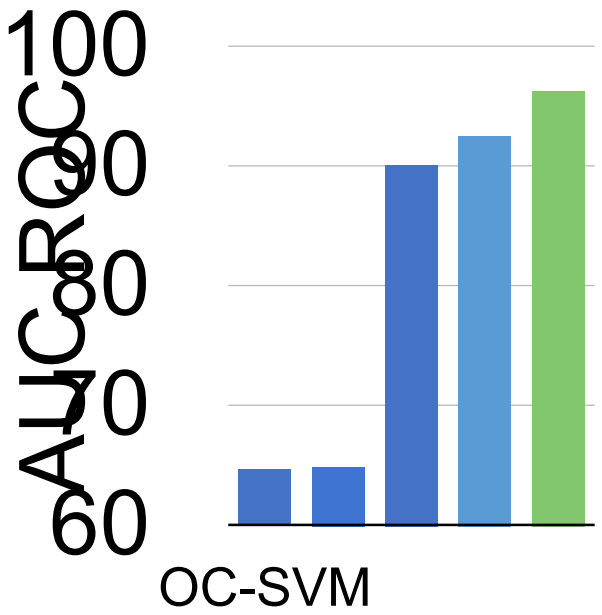
- Instead of combatting forgetting we embrace it!
- Fine-tune pre-trained features with the compactness criterion
- Anomaly detection accuracy initially improves, then degrades:



Seizing the moment

How to select the correct time to stop in unsupervised way?

- Simple solution: early stopping after constant number of iterations
  - Rely on a hyper-parameter that may not generalize to new datasets
- There are most principled solution but they don't work much better



Dataset	Self-Supervised			Pretrained	
	OC-SVM	DeepSVDD	MHRot	S.Baseline	PANDA
CIFAR10	64.7	64.8	90.1	92.5	**96.2**
CIFAR100	62.6	67.0	80.1	**94.1**	**94.1**
FMNIST	92.8	84.8	93.2	94.5	**95.6**
CatsVsDogs	51.7	50.5	86.0	96.0	**97.3**
DIOR	70.7	70.0	73.3	93.0	**94.3**

Our Latest Ideas: Mean Shifted and DINO

- Our recent work has improved over PANDA in different key aspects
- Moving away from using a compactness loss which assumes a singe normal class
- Instead using the contrastive objective which can accommodate many normal classes
-

Approach	Self-supervised		Pretrained		Hybrid		
Method	RotNet [14]	CSI [35]	ResNet	DINO	PANDA [27]	MSAD [28]	DINO-FT
CIFAR-10	90.1	94.3	92.5	97.1	96.2	97.2	**98.4**

Reiss and Hoshen. Mean-shifted contrastive loss for anomaly detection. *AAAI', 23*

Desideratum #2: Interpretable Video AD

- Deep learning is often not interpretable
- This is critical for AD – we never gave computer any examples of anomalies
- Here, we tackled task of video AD

Reiss and Hoshen, Attribute-based Representations for Accurate and Interpretable Video Anomaly Detection, Arxiv 2022

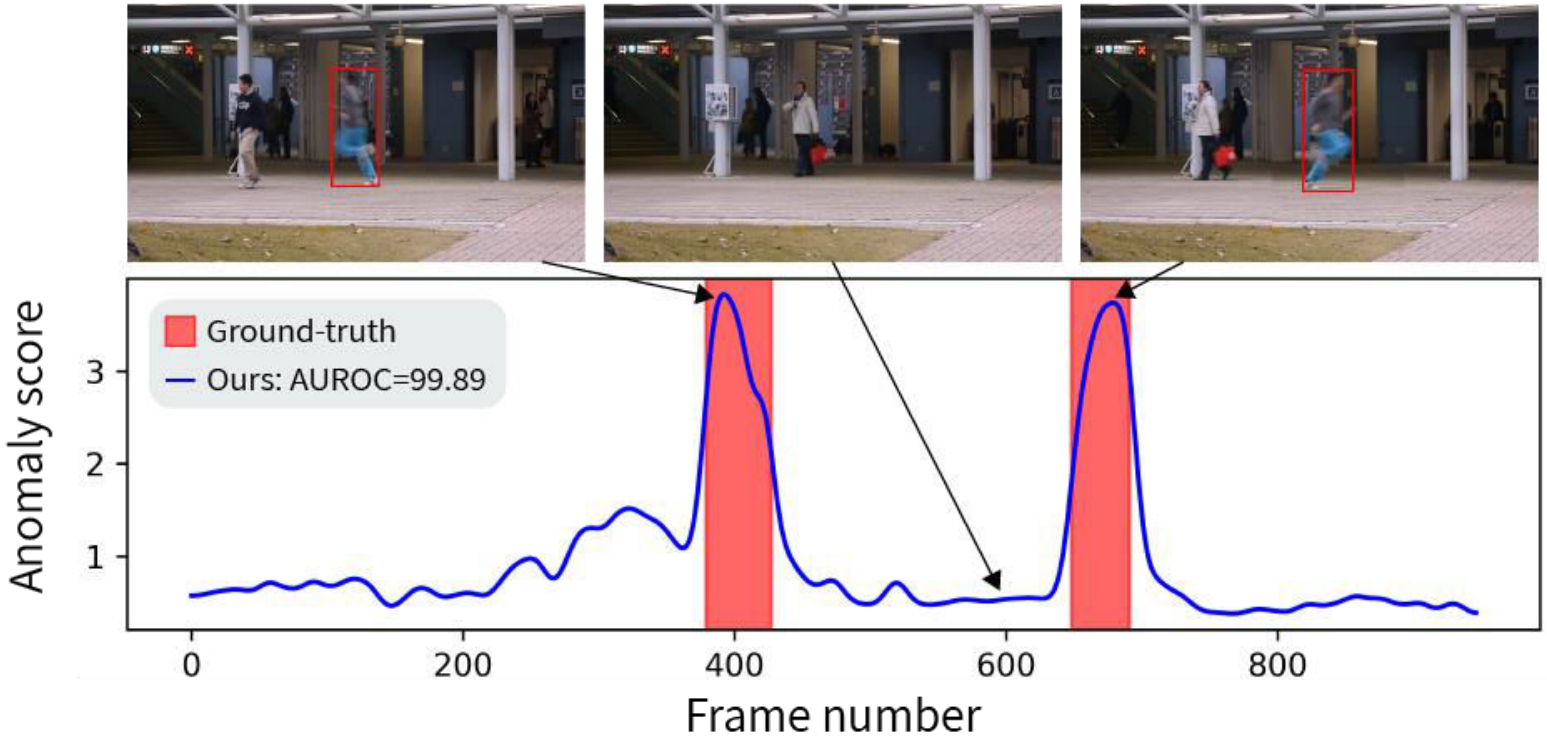
Unusual Velocity

- In current datasets, unusual activity is expressed by unusual velocity or human pose

Unusual Human Pose

Attribute-Based Representation

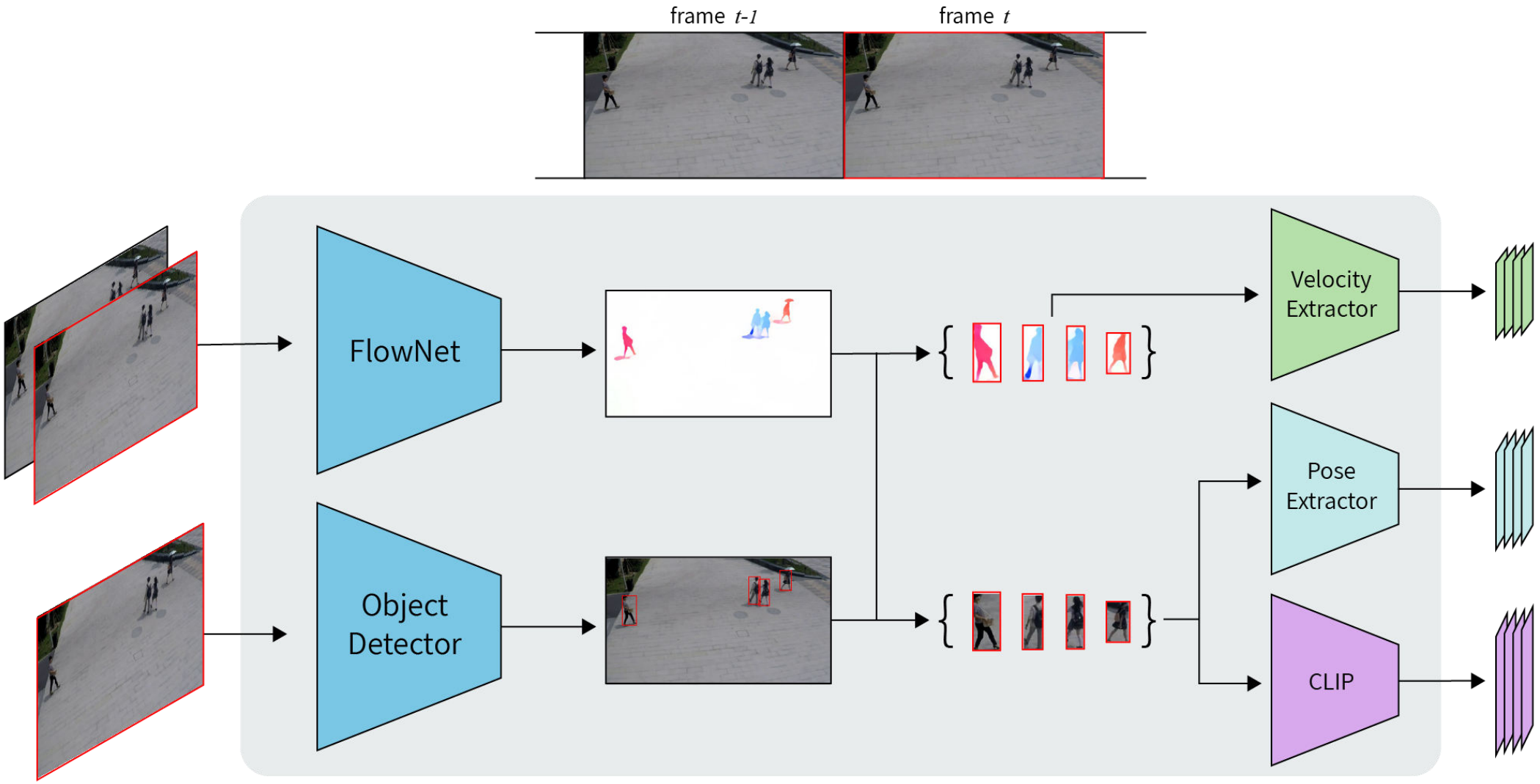
- We opt for an attribute-based representation for every object
- Velocity attribute: represented by an optical-flow histogram with 9 orientation bins
- Pose attribute: represented by the (x,y) coordinate of the human body landmarks
- Estimate density for each attribute separately
- This is inherently interpretable



Deep Residual

- Not every anomaly is due to one of the two attributes
- We detect residual attributes using a deep (non-interpretable) representation
- We use a CLIP image extractor, kNN density estimator

Entire System



Surprisingly, This is SoTA

- Our method outperform much more complex, and less interpretable methods
- Academic unsupervised video AD datasets are probably too easy

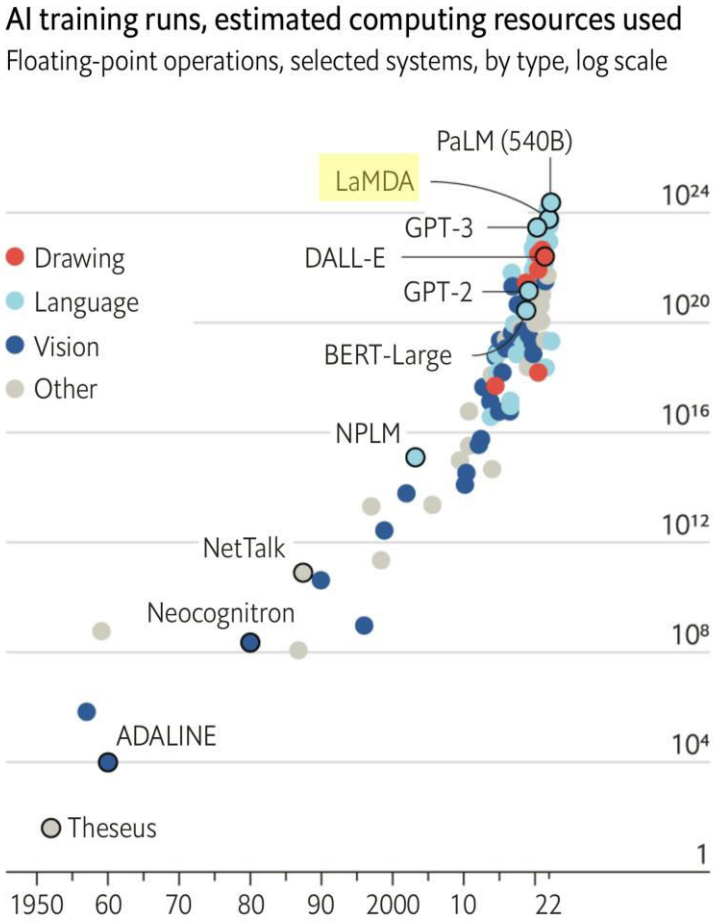
Year	Method	Ped2		Avenue		ShanghaiTech	
		Micro	Macro	Micro	Macro	Micro	Macro
2021	AMMCN [2]	96.6	-	86.6	-	73.7	-
	SSMTL [13]	97.5	99.8	91.5	91.9	82.4	89.3
	MPN [39]	96.9	-	89.5	-	73.8	-
	HF^2 [33]	**99.3**	-	91.1	<u>93.5</u>	76.2	-
	CT-D2GAN [12]	97.2	-	85.9	-	77.7	-
	BA-AED [14]	98.7	99.7	92.3	90.4	82.7	89.3
2022	BA-AED [14] + SSPCAB [53]	-	-	<u>92.9</u>	91.9	83.6	89.5
	DLAN-AC [62]	97.6	-	89.9	-	74.7	-
	Jigsaw-Puzzle [58]	99.0	**99.9**	92.2	93.0	<u>84.3</u>	**89.8**
	Ours	<u>99.1</u>	**99.9**	**93.3**	**96.2**	**85.9**	<u>89.6</u>

Desideratum #3: Dealing with Ambiguity

- Two humans can disagree on data deemed anomalous
- There exists a coordinate system to make every point anomalous
- How can ML deal with this ambiguity?

Can Scaling Up FM Improve AD Indefinitely?

- We saw that powerful FMs result in large improvements to AD
- Begs the question: can scaling-up even more solve AD?



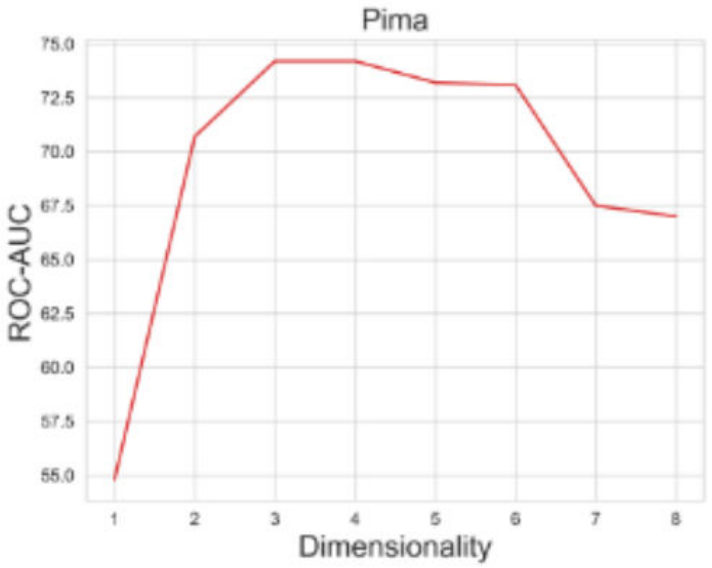
Computation vs. accuracy in language models

Example: Anomalous Bird Detection

- Training: many bird images
- Variation attributes: species, background, wing length, pose …
- Test:
  - Normal: same distribution as training
  - Anomaly: very long beak
- Can we use FMs to detection anomalies?

Expressivity-Sensitivity Tradeoff

- Assume perfect FMs, express all image attributes
- Representation: expressive enough to express **unspecified** anomalous attribute
- Increased expressiveness reduces sensitivity to anomalous attribute
- With Gaussian assumptions, sensitivity $\propto 1/\sqrt{n_{attributes}}$



Reiss, Cohen, Hoshen, **No Free Lunch: The Hazards of Over-Expressive Representations in Anomaly Detection**

No Free Lunch

Principle: A successful anomaly detection algorithm must choose the smallest number of attributes which include the (unspecified) anomalous attribute

Informal Collorary: there is no one FM representation that fits all

Asking Users for Guidance

- How can users guide the learned representations?
- A **few labeled** anomalies: will not generalize to unseen anomalies
- Ask about the relevant and nuisance attributes:
  - Relevant: ”Object category is the **most relevant** attribute for detecting anomalies”
  - Nuisance: “Gender is **irrelevant** for detecting anomalies”

Nuisance Guidance for AD: Example

- Nuisance attribute guidance: car model is irrelevant for anomaly detection
- Training set: side facing coupes and front facing sedans
- Front facing jeep: normal
- Back facing jeep: anomaly

x – image
d – car model
y – car pose

Representations Invariant to Nuisance Attributes

- Idea: ignore nuisance by learning representations invariant to nuisance attributes
- The invariant representation can then be used for kNN retrieval

[1] Kahana and Hoshen. A contrastive objective for learning disentangled representations. *ECCV'22* ,

[2] Cohen, Kahana and Hoshen, RedPANDA, Arxiv'22

Task

Every
image is labeled with the nuisance attribute
value: $(x_1, d_1), (x_2, d_2), .. (x_N, d_N)$

All other attributes are bundled into (unknown) attributes: $y_1, y_2, .. y_N$

Assumption: y and d are **independent**. d_i. y_i

Every image is labeled with the nuisance attribute
value: $(x_1, d_1), (x_2, d_2), .. (x_N, d_N)$
All other attributes are bundled into (unknown) attributes: $y_1, y_2, .. y_N$
Assumption: y and d are **independent**. d_i. y_i

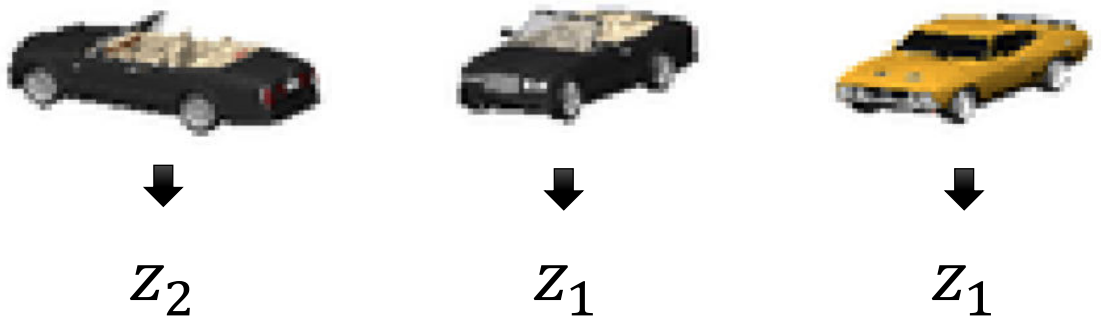
x – image
d – car model
y – car pose

Representation for Unknown Attribute

Sample	x_i	Image
Domain	d_i	Car model
Unknown attribute	y_i	Car pose
Learnt code	z_i	-

Learn an encoder E, where $z_i = E(x_i)$ so that:

- The representation contains all information about the relevant attributes: $z_i \Rightarrow y_i$
- The representations is invariant to the nuisance attribute: $z_i \quad d_i$



z_2 z_1 z_1

Sample	x_i	Image
Domain	d_i	Car model
Unknown attribute	y_i	Car pose
Learnt code	z_i	-

Disentanglement Goals

Two complementary metrics are required:

- Invariance – z does not reveal any information about the domain d

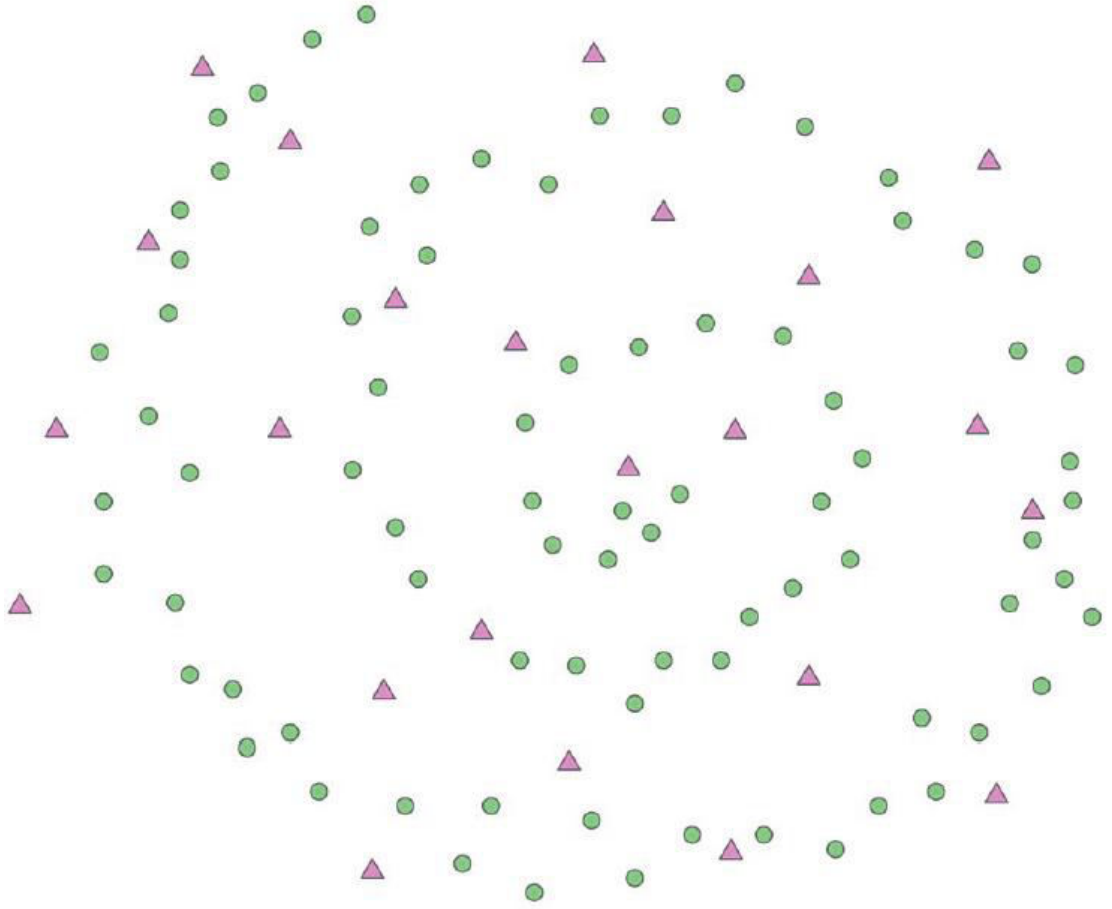
$$P(d|z) = P(d)$$

- Informativeness - z has all information about unknown attributes y

$$P(y|x) = P(y|z)$$

Theoretical Challenges for Anomaly Detection

- Characterizing desirable representations, under different priors for the distribution of anomalies.
- Characterizing conditions (distributional, supervision) when such representations be learned from the normal-only training data.



Practical Challenges for Anomaly Detection

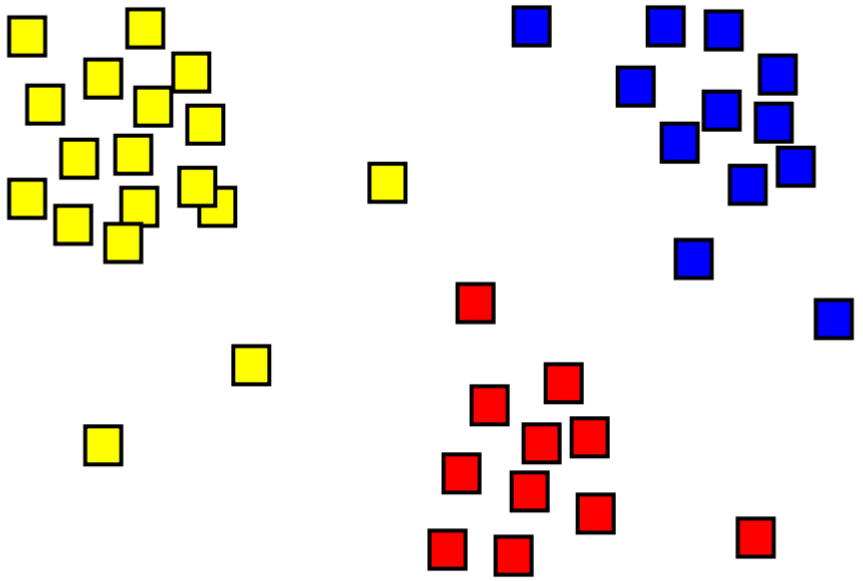
- Fine-grained AD: “Given all my previously seen birds, is this new bird anomalous?”
- Interpretable AD for more general settings
- Guidance: more effective human guidance with less effort
- Desidertaum 4: AD beyond images

Clustering

- Essential unsupervised task
- Classify a set of images **without** training labels

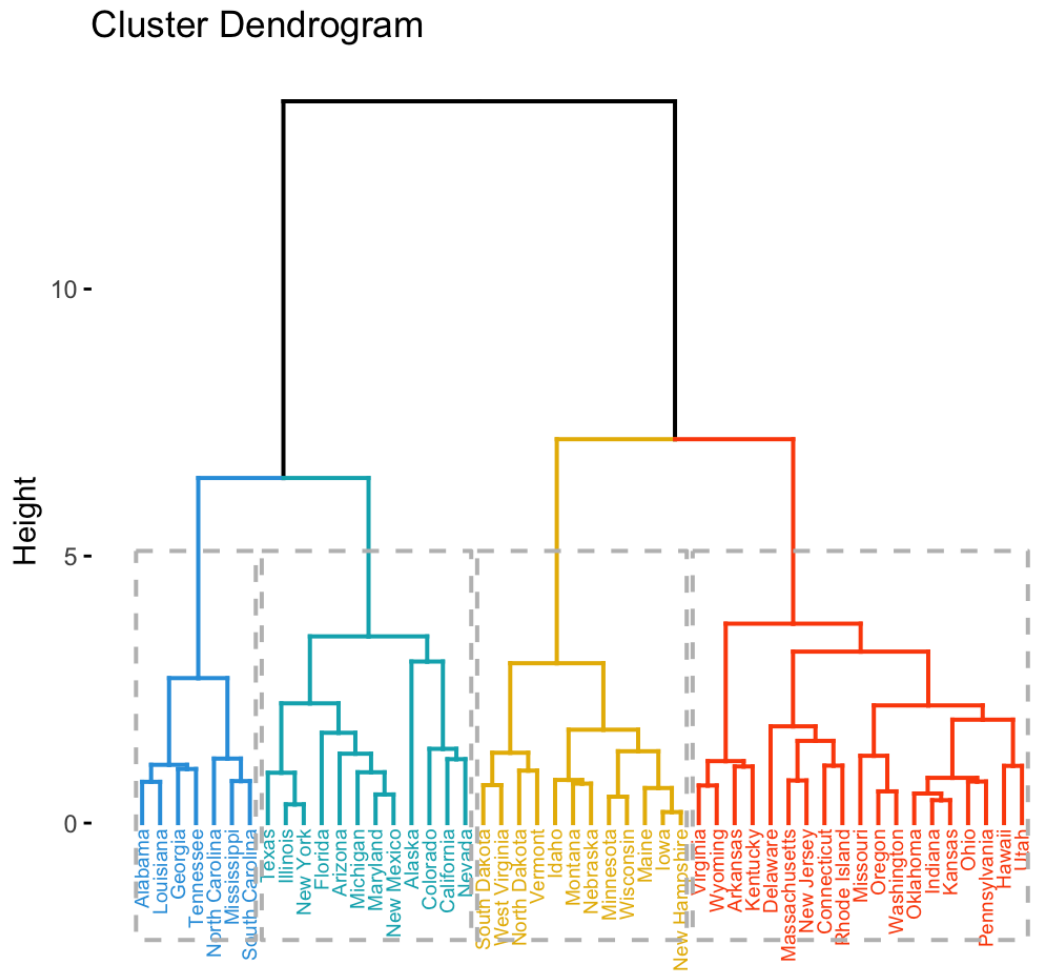
Is Unsupervised Classification Ever Possible?

- Hypothesis: classes form clusters in representation space
- By detecting clusters – discover classes



Hierarchical Clustering

- Every point is initially designated a cluster
- Merge clusters whose variance will increase the least
- Repeat until only K clusters remain



K Means Clustering

- You have seen this in IML

Algorithm 1 k-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

Latent-Variable Clustering

- This can be formulated as a **discrete** latent variable model
- Differently from VAE, here the latent variable has a small set of values

$$\log(p(x)) = \log(\sum_{z=1}^{K} p(x,z)) = \log(\sum_{z=1}^{K} p(x|z)p(z))$$

- This is easy to compute
- No need for importance sampling and proposal distributions

Classification With Latent Variable Models

- To classify using latent variable model, use Bayes' law

$$p(z = i|x) = \frac{p(x|z{=}i)p(z{=}i)}{p(x)} = \frac{p(x|z{=}i)p(z{=}i)}{\sum_k p(x|z{=}k)p(z{=}k)}$$

Gaussian Mixture Models (GMMs)

- In GMMs, we assume that p(x|z) is Gaussian

$$p(x|z) = N(\mu_z, \Sigma_Z) \quad p(z) = s_z$$

- All parameters are optimized using maximum likelihood

$$max_{\mu_1..\mu_K, \Sigma_1..s_1..s_K} \sum_{i=1}^{N} \log(p(x_i))$$

- Extend to deep encodings by learning E and clustering $e_i = E(x_i)$

Classification-Based Clustering

- We make two assumptions:
  - We know the proportion of samples in each cluster (e.g. equal)
  - Every sample belongs to a single cluster only

Distribution Loss

- Compute q(z=i) proportion of data assigned to z=i:

$$q_\theta(i) = E_x p(z = i|x)$$

- Loss is therefore the difference from the prior:

$$L_{prop} = KL(q_\theta||r)$$

Classifier Confidence

- We want classifiers to assign point confidently to a single class
- Enforce low entropy, which does exactly that

$$L_{confidence} = -E_x \sum_i p_\theta(z = i|x) \log(p_\theta(z = i|x))$$

Optimize Loss with Some Representation Loss

- Optimize both confidence and proprtion loss
- Typically easy to overfit, therefore all include some rep learning loss

$$\min_\theta L_{prop} + L_{confidence} + L_{contrastive}$$

Domain Generalization

- Assume we have a labeled training set for domain A (photos)
  - A dataset with pairs of $(x_i, y_i) \in D_A$
- We want a classifier that performs well on domain B (paintings)
- Can we make. a classifier that generalizes cross domain?

Domain Generalization – Nothing Beats ERM

- Nothing beats empirical risk minimization (just training) on domain A
- As always, powerful pretraining and regularization helps
- Lots of celever things tried, nothing really works ☹

In Search of Lost Domain Generalization

Ishaan Gulrajani and David Lopez-Paz*
Facebook AI Research
igul222@gmail.com, dlp@fb.com

Domain Adaptation

- Assume we have a labelled training set from domain A
- We also have an **unlablled** set of images from domain B
- Can we now do better?

Combining Classification + Clustering

- We can combine classification loss for A and clustering losses for B

$$E_{x \in D_A} KL(p_\theta(x) || 1_{y_i}) + E_{x \in D_B} KL(q_\theta || r_B) + E_{x \in D_B} H(p_\theta(z|x))$$

- Limitation: works when classifier trained on A works ok on B
- Otherwise, even if clustering works, clusters will be mislabeled

Looking Back

- In this course, we focused on long standing ML ideas
- Modern breakthroughs are simple (but clever) extension

Looking Back

- In this course, we focused on long standing ML ideas
- Modern breakthroughs are simple (but clever) extension

Generative models – AR Models

- Classical: Auto-regressive models on discrete data
- SoTA: Replace the function approximator by a transformer

Generative models – Variational Inference

- Classical: Variational approximation for likelihood estimation of latent variable models using importance sampling
- SoTA: Replace the latent optimization of thr sampling distribution $q(z|x) = N(mu\_x, sigma^2\_x)$ by an amortized encoder E(x)

Generative models – Diffusion models

- Classical idea:
  - 1) create a forward diffusion model by adding Gaussian noise to sample
  - 2) invert this process by the reverse diffusion process, requiring the score
- Idea for 2010 (Vincent) – score estimation by denoising
- SoTA idea: use deep network for denoising

Generative Models – IPM and GANs

- Classical idea: compute distance between distributions using the integral probability metric (IPM) with Lipschitz-1 functions
- SoTA idea:
  - 1) Use to train generator networks minimizing distance between gen and real
  - 2) Estimate IPM using a deep neural network discriminator

Representation Learning – Local and VicReg

- Classical ideas:
  1. Compute local region by finding kNN using L2
  2. Enforce embeddings of local regions to be similar
- SoTA idea:
  - Use augmentations to create local region
  - Use encoder to map sample to representation

Rep. Learning – SNE and Contrastive

- Classical ideas:
  1. Compute distribution-like weighting between points using softmax on L2
  2. Do the same thing for embeddings
  3. Optimize eembeddings using KL between distributions
- SoTA idea:
  - Use augmentations to create one-hot distribution for points
  - Use encoder to map sample to embedding

Rep. Learning – CCA and Barlow Twins

- Classical ideas:
  1. Embeddings of two views of the same sample have correlation identity
- SoTA idea:
  - Use augmentations to create views
  - Use encoder to map sample to representation