Early Detection of Backbone Divergence via Neighborhood Structure in Learned Embeddings

Yiming Zhang[1,2]
Reza Zare Hassanpour[1] Georgi Gaydadjiev[3]

[1] Bernoulli Institute, University of Groningen
[2] CogniGron, University of Groningen
[3] Delft University of Technology

NeurIPS 2026
9–11 December, 2026
Paris, France

COGNIGRON

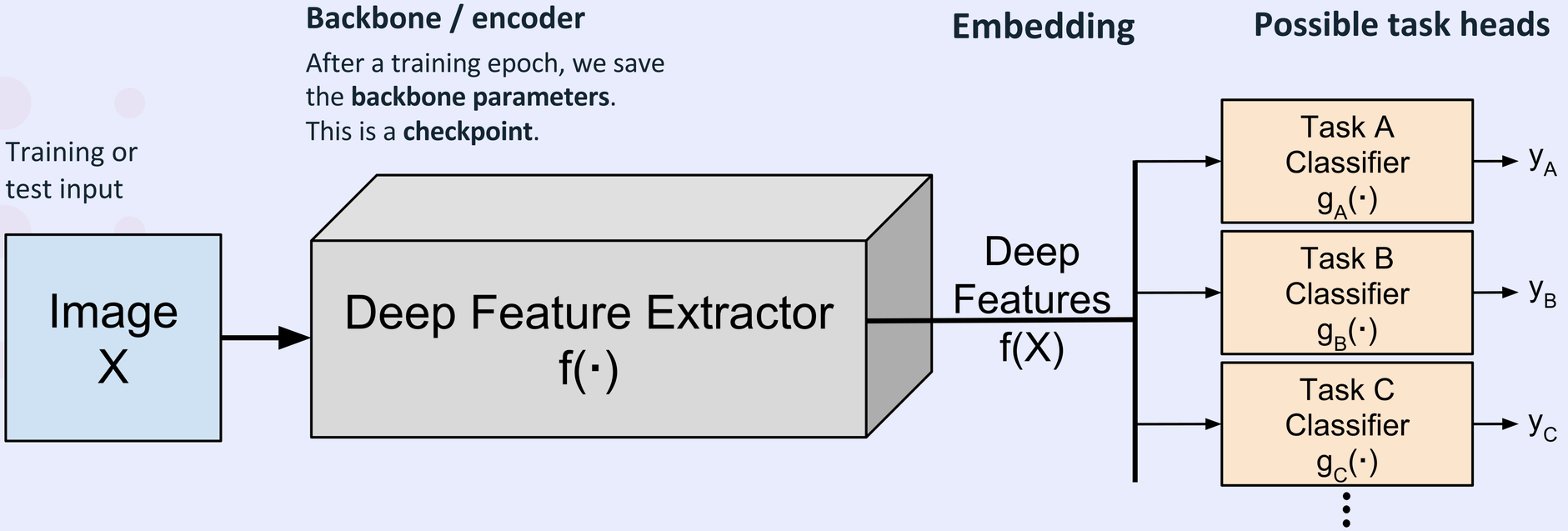
NEURAL INFORMATION PROCESSING SYSTEMS

university of groningen

UEF Ubbo Emmius Foundation
Moving science forward.

When do train and test embeddings start to diverge?

The backbone maps each input to one learned feature vector, called an embedding.
A task head uses the embedding for downstream tasks, such as prediction.



Motivational example: Image Classification

Loss and accuracy do not directly compare training and test embedding distributions.

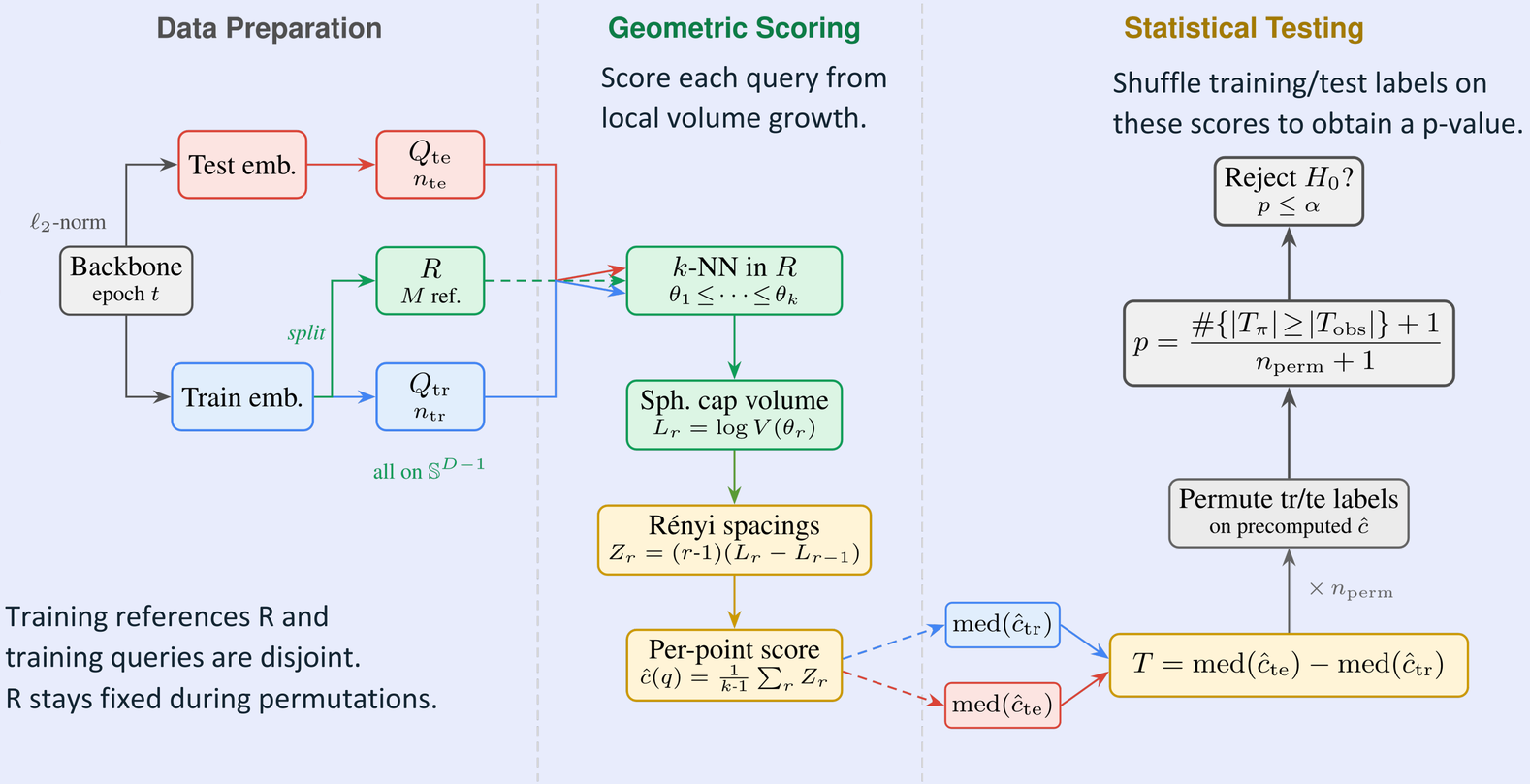
At which checkpoints can we detect train–test embedding differences?

Figure: Mayer, Bayar & Stamm, Learning Unified Deep-Features for Multiple Forensic Tasks, IH&MMSec 2018, Fig. 1 (cropped).

Our proposed Volume-Ratio Test (VRT)

VRT compares local neighbourhood structure around training and test embeddings.

H_0: training and test embeddings follow the same distribution.



Training references R and training queries are disjoint.
R stays fixed during permutations.

p < 0.05 → Detect a distribution difference

When do VRT and the baselines detect train–test differences?

Shown: **17 of 24** models used to compare detection timing.
The other seven cover regularization experiments and negative controls (paper, Table 1).

Detection onset (epoch, *lower means earlier*). **Normalized embeddings, eval-mode extraction**.

Onset = first checkpoint with detection continuing at every later evaluated checkpoint.

Model / data	VRT	MMD	ED	MMDAgg	C2ST
ResNet20-FRN / CIFAR-10	**38**	—	—	—	—
ResNet-18 / CIFAR-10 (×5)	**25–30**	—	—	40–42	—
ResNet-18 / CIFAR-100 (×5)	**26–34**	43–45	43–45	44–46	—
ResNet-50 / CIFAR-100	**61**	73	73	80	121
Phase Zoo / CIFAR-100 (batch size 8)	**150**	150	150	150	150
Phase Zoo / CIFAR-100 (batch size 128)	**≤100**	≤100	≤100	≤100	≤100
Phase Zoo / CIFAR-100 (batch size 1,024)	**≤100**	≤100	≤100	≤100	125
SimCLR / CIFAR-10	**229**	284	283	483	691
SimCLR / CIFAR-100	**191**	213	213	348	653

—: no sustained detection. ≤100: detected at the first available checkpoint.
Ranges summarize five independently trained models in each zoo.
MMD: maximum mean discrepancy
ED: energy distance
MMDAgg: MMD aggregated two-sample test
C2ST: classifier two-sample test

VRT detects first or ties for first across all 17 models shown.

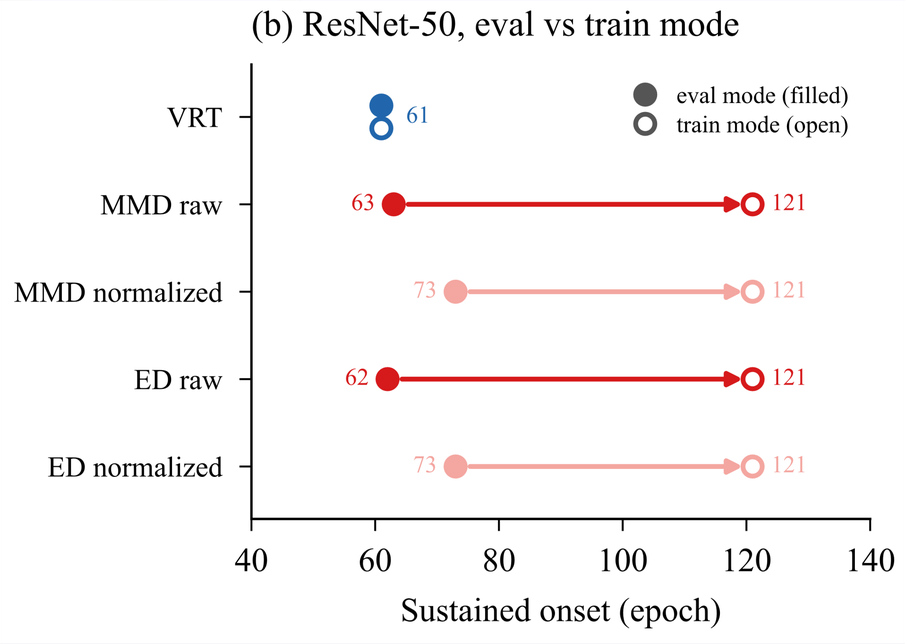
How does BatchNorm mode affect detection?

Model weights stay fixed. Only the BatchNorm extraction mode changes.
All 16 BatchNorm models from the main comparison (paper, Table 2).

Eval mode Running statistics from training

Train mode Statistics from the current batch

Example: ResNet-50



Other 15 BatchNorm models

Normalized embeddings · onset epoch, **eval → train**

Model	VRT	MMD	ED
ResNet-18 C100 zoo ×5	**26–34 → 20–26**	43–45 → 55–56	43–45 → 54–55
ResNet-18 C10 zoo ×5	**25–30 → 26–32**	—	—
Phase Zoo (batch size 8)	**150 → 150**	150 → 150	150 → 150
Phase Zoo (batch size 128)	**≤100 → ≤100**	≤100 → ≤100	≤100 → ≤100
Phase Zoo (batch size 1,024)	**≤100 → ≤100**	≤100 → 125	≤100 → 125
SimCLR CIFAR-10	**229 → 527**	284 → 692	283 → 687
SimCLR CIFAR-100	**191 → 324**	213 → 583	213 → 583

—: **no sustained detection** in either mode; ranges span five zoo models.
≤100: detection at the first evaluated checkpoint.

Changing BatchNorm mode can delay MMD and ED much more than VRT.
VRT still detects earlier or at the same time.

What changes when MMD uses VRT scores?

Same MMD test, different input: normalized embeddings or their neighbourhood scores ĉ.

Each score ĉ describes **local volume growth** around one embedding.

Model / data	MMD on normalized embeddings	MMD on neighborhood scores ĉ	Change
ResNet20-FRN / CIFAR-10	—	**24**	Detection appears
ResNet-18 zoo / CIFAR-10	—	**23–28**	Detection appears
ResNet-50 / CIFAR-100	73	**61**	12 epochs earlier
ResNet-18 zoo / CIFAR-100	43–45	**23–29**	Earlier in all 5 models
SimCLR / CIFAR-10	284	**234**	50 epochs earlier
SimCLR / CIFAR-100	213	**187**	26 epochs earlier

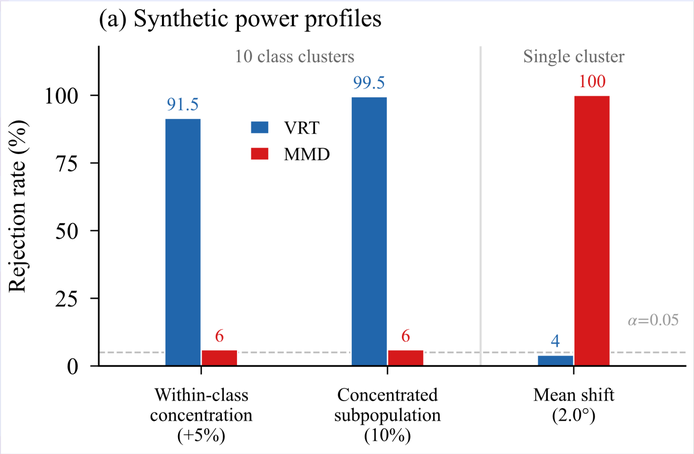
Eval-mode extraction. Detection onset (epoch). —: **no sustained detection**.
Ranges span five zoo models. Appendix I, Figure 9.

VRT scores help MMD detect earlier, including cases with no detection before.

Analysis: Why does VRT detect earlier?

Which embedding changes do VRT and MMD detect?

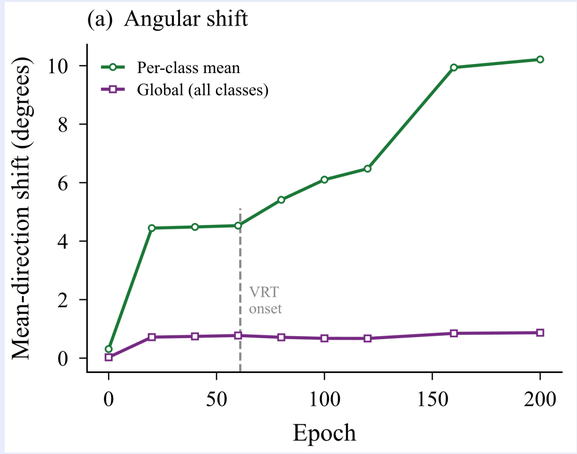
Synthetic data · 200 runs per case



VRT detects **tighter clusters** more often.
MMD is more sensitive to the **direction shift** in this comparison.

Class-level versus overall mean shifts

ResNet-50 / CIFAR-100 · Appendix J



Green: compare each class, then average the angles.
Purple: combine classes, then compare overall centres.

MMD detects an overall direction shift well in the synthetic example.
In the real embeddings, class shifts can **partly cancel** when we combine classes, leaving a **small overall shift** that MMD may not detect.

Local comparisons can reveal differences even when the overall direction changes little.

In conclusion:

VRT detects train–test differences earlier in many settings, including cases where other tests fail to detect them.

Future work could explore when train–test differences become detectable in word or sentence embeddings from language models.

Paper

Code and data

yiming.zhang@rug.nl

university of groningen / faculty of science and engineering / bernoulli institute

The authors would like to acknowledge the financial support of the CogniGron research center and the Ubbo Emmius Funds (Univ. of Groningen).