# A Representation Analysis of Image Anomaly Detection

Tom Dietterich

Work by: Risheek Garrepalli (now at QualComm), Alex Guyer

Advice from: Alan Fern, Fuxin Li, Stefan Lee

Oregon State University

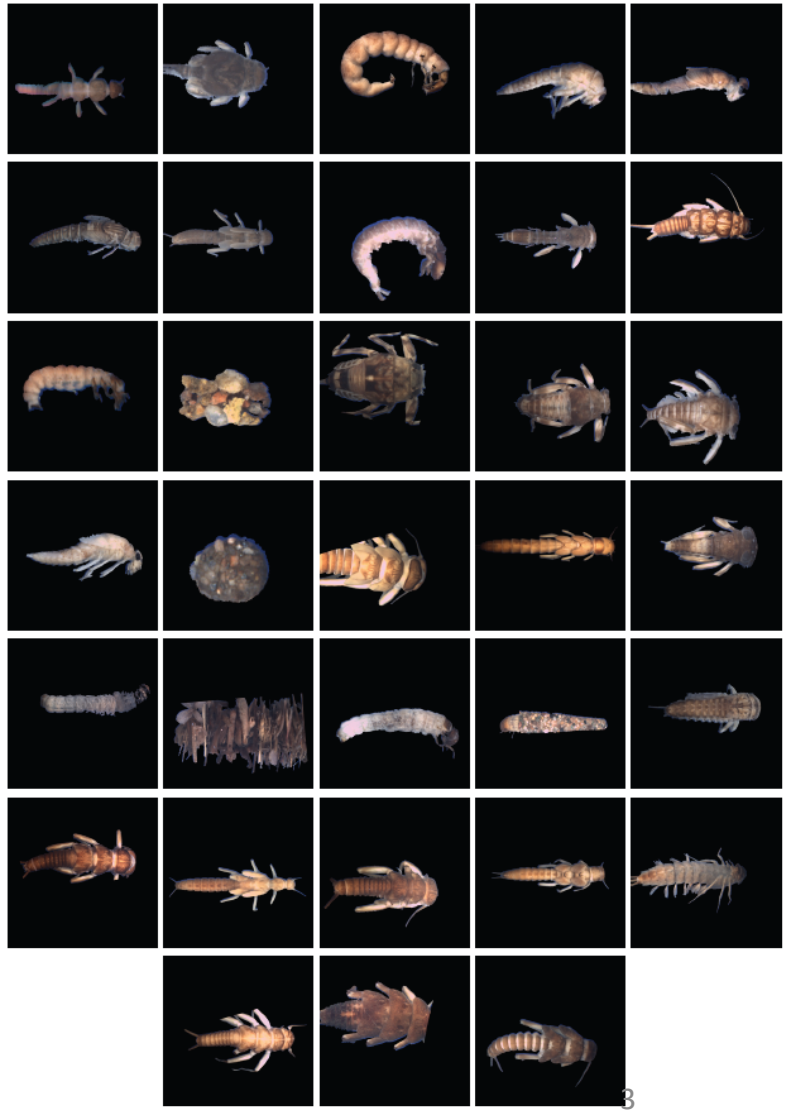
# Motivation: Automated Counting of Freshwater Macroinvertebrates

- Goal: Assess the health of freshwater streams
- Method:
  - Collect specimens via kicknet
  - Photograph in the lab
  - Classify to genus and species
- BugID Project
  - 54 classes of interest to the EPA
  - accuracy $\approx 90\%$
  - Larios, N., Soran, B., Shapiro, L., Martínez-Muños, G., Lin, J., Dietterich, T. G. (2010). **Haar Random Forest Features and SVM Spatial Matching Kernel for Stonefly Species Identification.** *IEEE International Conference on Pattern Recognition (ICPR-2010)*.
  - Lin, J., Larios, N., Lytle, D., Moldenke, A., Paasch, R., Shapiro, L., Todorovic, S., Dietterich, T. (2011). **Fine-Grained Recognition for Arthropod Field Surveys: Three Image Collections**. *First Workshop on Fine-Grained Visual Categorization (CVPR-2011)*
  - Lytle, D. A., Martínez-Muñoz, G., Zhang, W., Larios, N., Shapiro, L., Paasch, R., Moldenke, A., Mortensen, E. A., Todorovic, S., Dietterich, T. G. (2010). **Automated processing and identification of benthic invertebrate samples.** *Journal of the North American Benthological Society*, 29(3), 867-874.

www.epa.gov

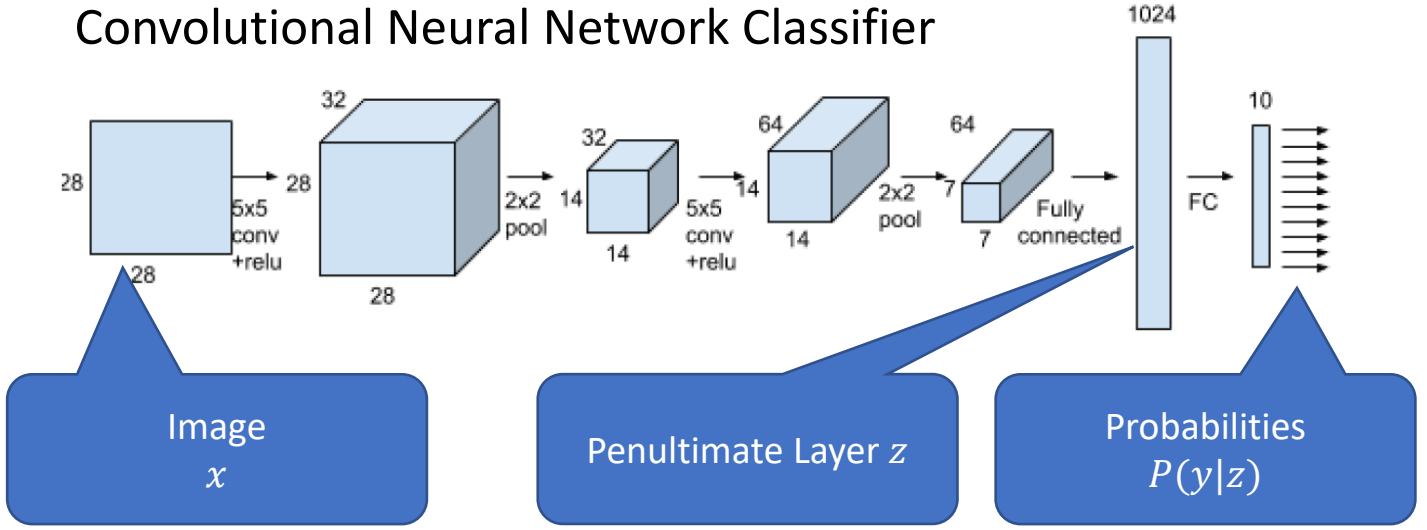
# Problem: There are $\approx \mathbf{76,000}$ species of freshwater insects worldwide

- 1200 species in US
- Field samples may contain other things
  - small rocks
  - leaves
  - trash
- Simple estimate of equal error rate for novel classes vs. the 54 classes was 20% (in 2011)
  - classifier is not usable without addressing the novel class problem
- We still need to solve this problem



# Baseline Method: Classifier Indecision

- **Classifier approach**
- Learn classifier $f(x) = P(y|x)$
- Compute a measure of uncertainty:
  - $A(x) = 1 - \arg\max_y P(y|x_q)$
  - $A(x) = H\left(P(y|x_q)\right)$
  - $A(x) =$ max class "logit"
- This should not work, because the classifier should discard all aspects of $x$ that are irrelevant to classification
- Surprise: it works fairly well
  - Hendrycks & Gimpel (ICLR 2017) "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks"




How much information does the latent space $z$ contain for anomaly detection?

# Research Questions

Risheek Garrepalli MS Thesis (2020)

- Q1: How well do existing anomaly scoring methods extract the anomaly information that is captured in the latent representation $z$?
  - Approach: Compare to an oracle anomaly detector
- Q2: How well could *any* network with this architecture perform the anomaly detection task?
  - Approach: Supervised training on both nominal and anomalous classes

- Definition of anomalies: Classes not seen during training
  - “Open Category” or “Open Set” problem
  - We claim this is harder and more realistic than Out-Of-Distribution tasks

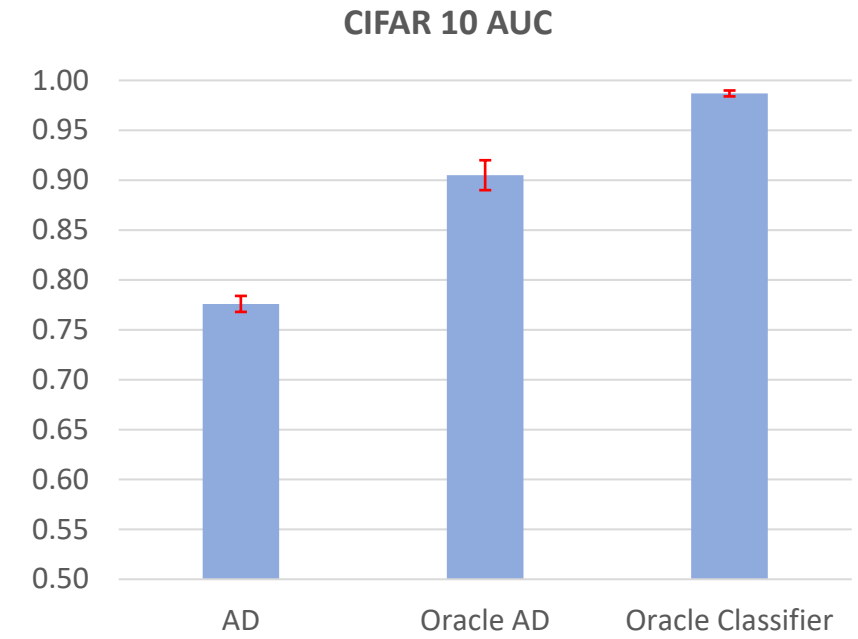
# Methods:

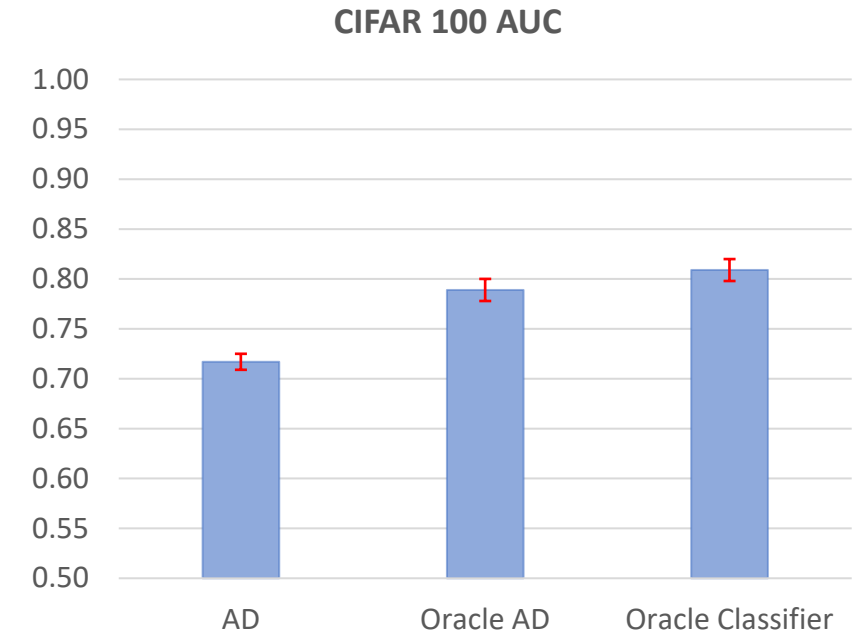
- CIFAR-10: 6 “nominal” classes and 4 “anomaly” classes
- CIFAR-100: 80 “nominal” classes and 20 “anomaly” classes

- Train Classifier
  - Divide data into train (60%), validate (20%), test (20%)
  - Remove anomaly classes from the training and validation data
  - Train ResNet34; use validation set accuracy to determine stopping point
  - Compute anomaly score on test set; measure AUC (“nominal” vs “anomaly” decision)
- Oracle Anomaly Detection
  - Take all validation data and label the nominal classes as “nominal” and the anomaly classes as “anomaly”
  - Train a random forest (1000 trees) that takes $z$ as input and predicts “nominal” vs. “anomaly”
  - Compute test set anomaly scores using this classifier; measure AUC
- Oracle Representation
  - Train ResNet34 on all classes
  - Train a random forest (1000 trees) that takes $z$ as input and predicts “nominal” vs. “anomaly”
  - Compute test set anomaly scores using this classifier; measure AUC

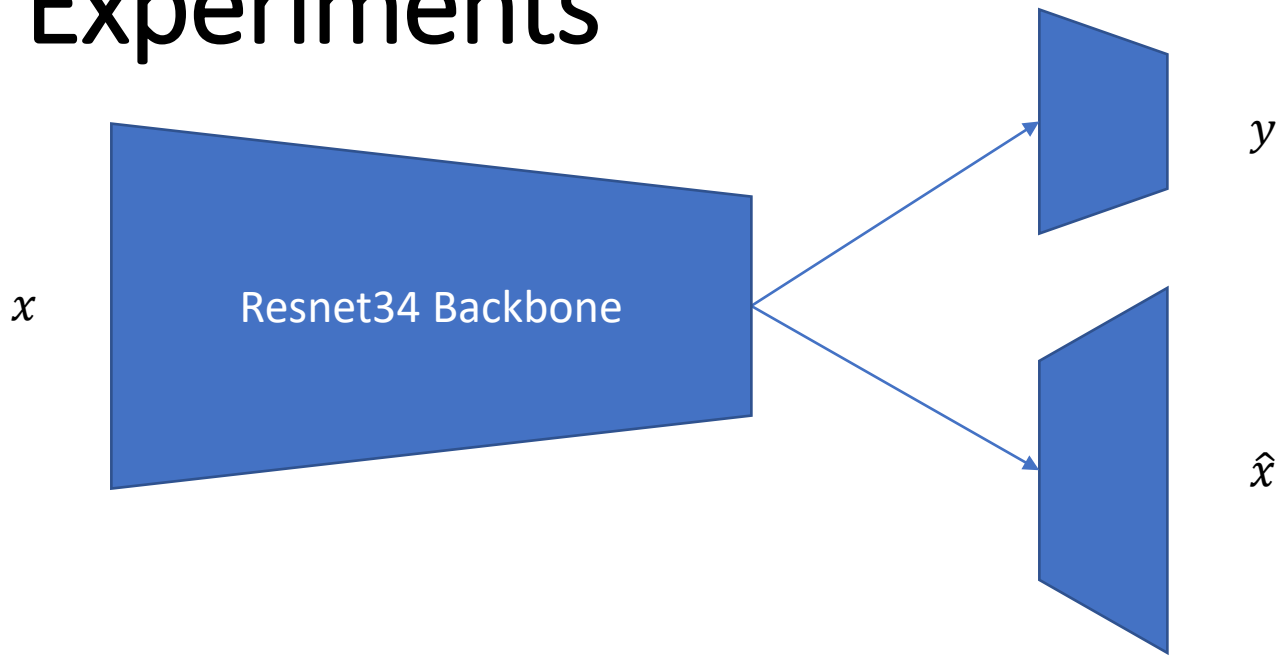
# Results

- Details:
  - Oracle Anomaly Detector: 1000-tree Random Forest
  - Anomaly Score: max logit
- Q1: The latent space contains much more anomaly information than is extracted by current anomaly scores
  - 0.776→0.905 = 0.129
  - 0.717→0.789 = 0.072
- Q2: There is additional anomaly information in the images that is not represented by the latent space
  - 0.905→0.987 = 0.082
  - 0.789→0.809 = 0.020



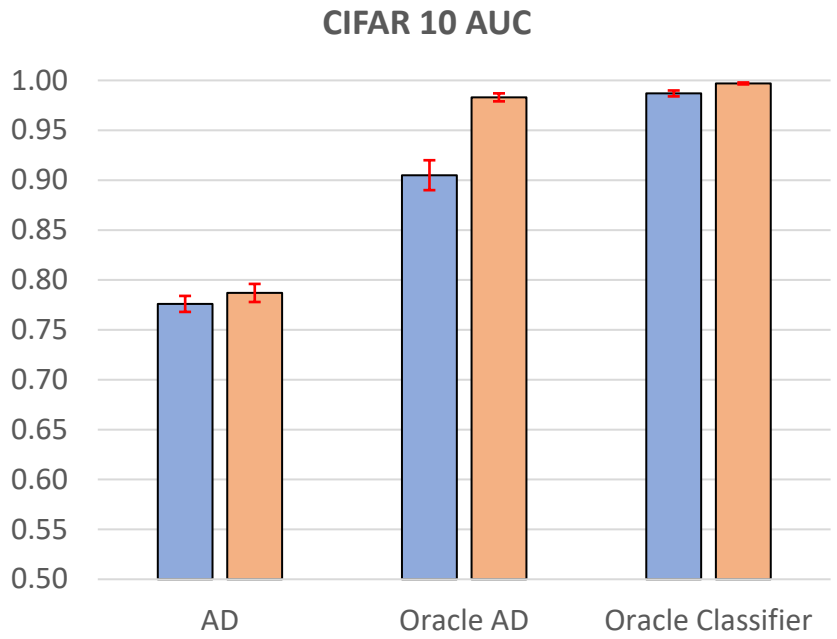
# Hybrid Network Experiments

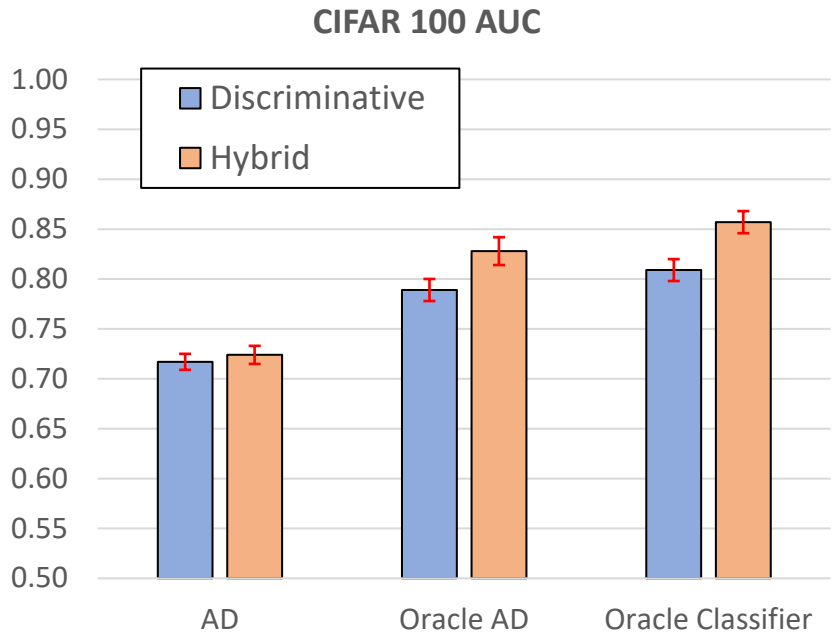



- Add a Reconstruction head to the network and jointly train the representation to support both classification and reconstruction (per-pixel squared error)
- Loss = Cross-Entropy + $\lambda \times$ Reconstruction Error
- CIFAR10: $\lambda = 0.9$, CIFAR100: $\lambda = 0.005$
- See also:
  - Oza, P., & Patel, V. M. C2AE: Class Conditioned Auto-Encoder for Open-set Recognition. CVPR 2019
  - Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., & Lakshminarayanan, B. (2019). Hybrid models with deep and invertible features. ICML 2019
  - Zhang, H., Li, A., Guo, J., & Guo, Y. Hybrid Models for Open Set Recognition. ECCV 2020
  - Perera, P., Morariu, V. I., Jain, R., Manjunatha, V., Wigington, C., Ordonez, V., & Patel, V. M. Generative-discriminative feature representations for open-set recognition. CVPR 2020

# Hybrid Network Experiments

- Q3: Does this increase the amount of anomaly-relevant information in the latent representation?
- Result: Hybrid representation improves performance
- Caution: $\lambda$ tuned using labeled test data



# Reflections

- Several people have observed that better classifiers allow simple anomaly detection methods (such as max logit) to work better
  - We have no theory why. Why doesn't the representation become more specialized to the prediction task and lose novelty detection power?
  - Are we just getting lucky with over-parameterized networks, because gradients $\rightarrow 0$ as the supervised loss $\rightarrow 0$?
- The reconstruction objective is theoretically sensible, but very difficult to train
  - Pure reconstruction models are *very* difficult to train
- The hybrid method works better, but it is not obvious how to tune $\lambda$
  - Why do we need the supervised head? Is it just regularizing the reconstruction representation?
- Some input augmentations improve performance of both classification and anomaly detection

# Improvements

[Alex Guyer]

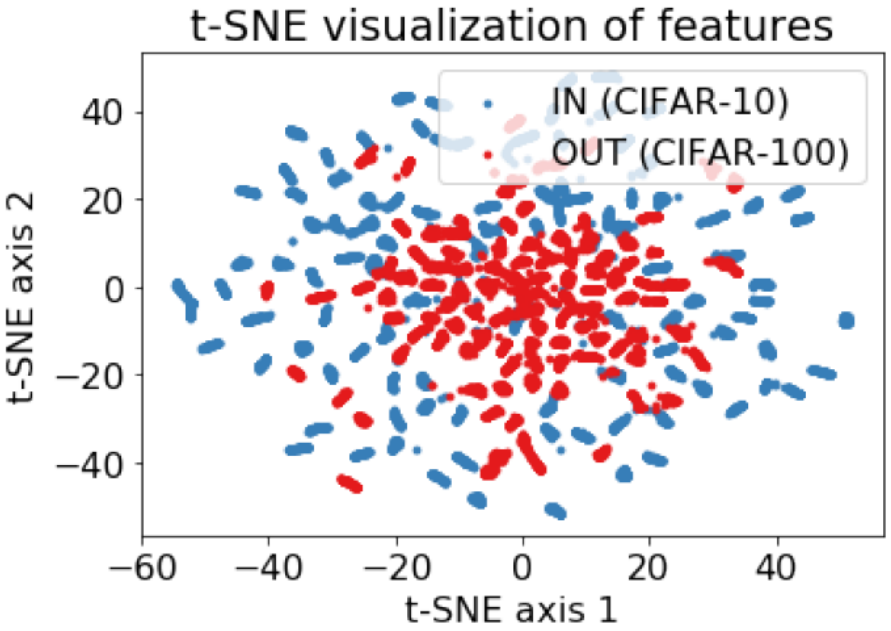
- Generalized ODIN
  - [Hsu, Y.-C., Shen, Y., Jin, H., & Kira, Z. (2020). Generalized ODIN: Detecting Out-of-distribution Image without Learning from Out-of-distribution Data. *CVPR 2020*]
  - $\mathrm{SoftMax}\left(\frac{h_1(z)}{g(z)}, \frac{h_2(z)}{g(z)}, \ldots, \frac{h_K(z)}{g(z)}\right)$ with $g$ and $h$ parameterized separately
    - We use cosine similarity for $h$ and linear model for $g$
  - The input perturbations from ODIN were not useful
  - Temperature scaling was not useful
  - Consistent but modest improvement (e.g., 1 percentage point on CIFAR100)
- Deep Ensembles and Deep SVGD Ensembles

# Some Other Things We Tried

[Alex Guyer]

- GEOM and other self-supervised tasks
  - [Golan, I., & El-Yaniv, R. Deep Anomaly Detection Using Geometric Transformations, *NeurIPS 2018*; Gidaris, S., Singh, P., & Komodakis, N. Unsupervised representation learning by predicting image rotations. *ICLR 2018.*]
  - rotation
  - recoloring
  - Sobel edge detection
  - Did not match baselines
- CSI (self-supervised representation)
  - [Tack, J., Mo, S., Jeong, J., & Shin, J. (2020). CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. *NeurIPS 2020*.]
  - Very sensitive to hyperparameters
  - Instance-level self-supervision is slow to train
  - Shows some promise
- JEM
  - [Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Swersky, K., & Norouzi, M. (2020). Your classifier is secretly an energy based model and you should treat it like one. *ICLR 2020*, 1–23.]
  - SVGD training is very slo

CSI feature visualization



# Concluding Remarks

- Latent representations from supervised learning contain information that we are not extracting
  - The information is not easily available (e.g., to linear models)
  - The information may not reflect open space or support few-shot learning
- Methods based on auxiliary or reconstruction tasks are theoretically sound but difficult to train in practice
- Hybrid methods show promise