# Automated Classification of Model Errors on ImageNet


Momchil Peychev*

Mark Niklas Müller*

Marc Fischer

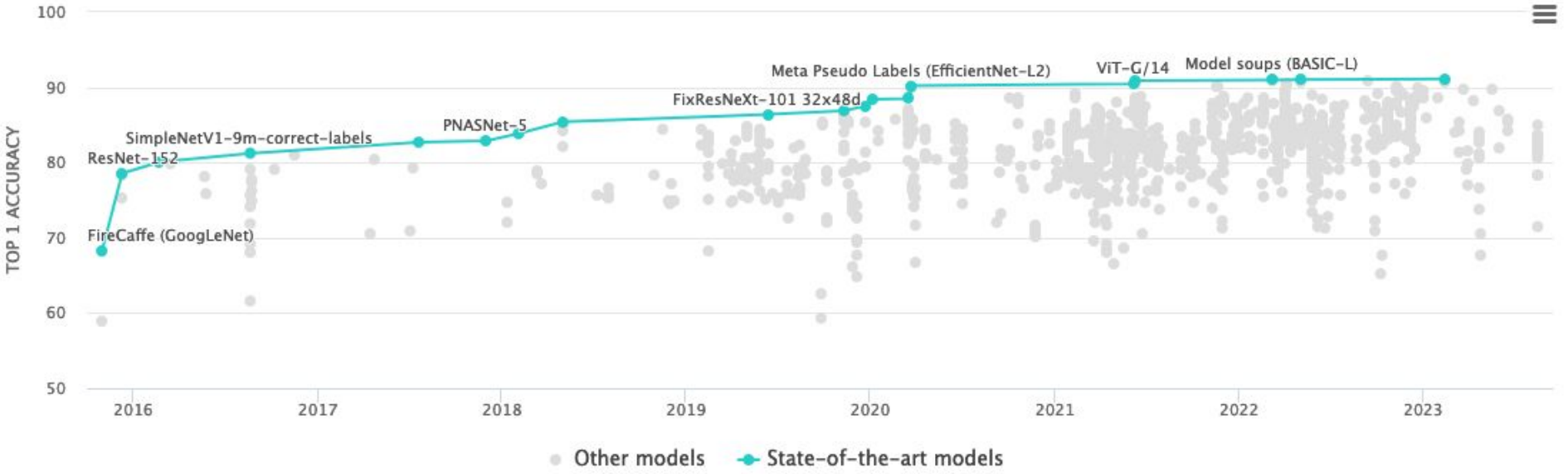
Martin Vechev

ETH zürich

NEURAL INFORMATION PROCESSING SYSTEMS

SRILAB

# ImageNet Progress



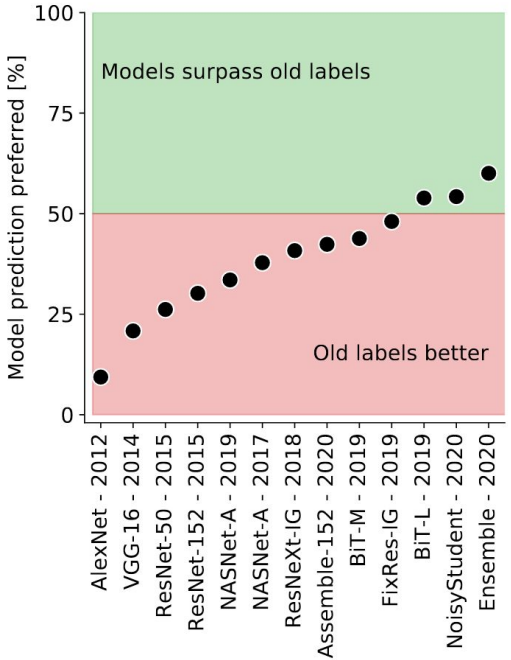

Source: [Papers with Code | Image Classification on ImageNet](#) (9 Nov 2023)

ImageNet still drives progress to date, but top-1 accuracy is stagnating.

# Model Predictions vs Ground-Truth

Humans prefer model predictions over the original labels.

*How can we further evaluate progress on ImageNet?*




(Figure from Beyer et al.)

Beyer et al., ”Are we done with ImageNet?”, arXiv 2020
Tsipras et al., “From ImageNet to Image Classification: Contextualizing Progress on Benchmarks”, ICML 2020

# Categorization of Model Errors on ImageNet

Prior work (Vasudevan et al.):

- **Manual review** by a panel of experts
- Classify **error category** and **severity**

✘ time-consuming
✘ inconsistent
✘ infeasible without experts

⇒ restricted to **two SOTA models**

Vasudevan et al., “When does dough become a bagel? Analyzing the remaining mistakes on ImageNet”, NeurIPS 2022

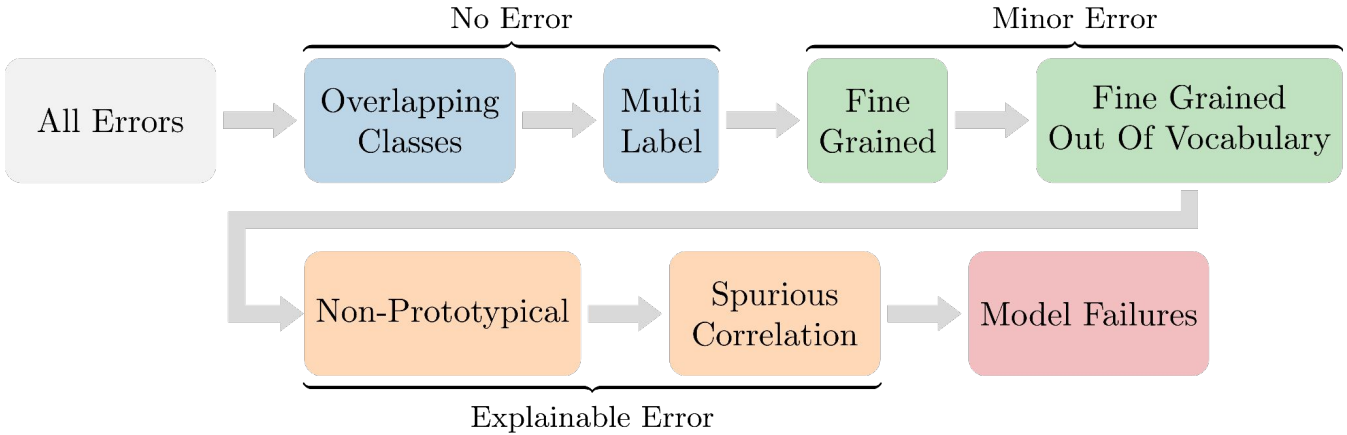
# **Automated** Classification of Model Errors

**This work:** Automated error classification pipeline

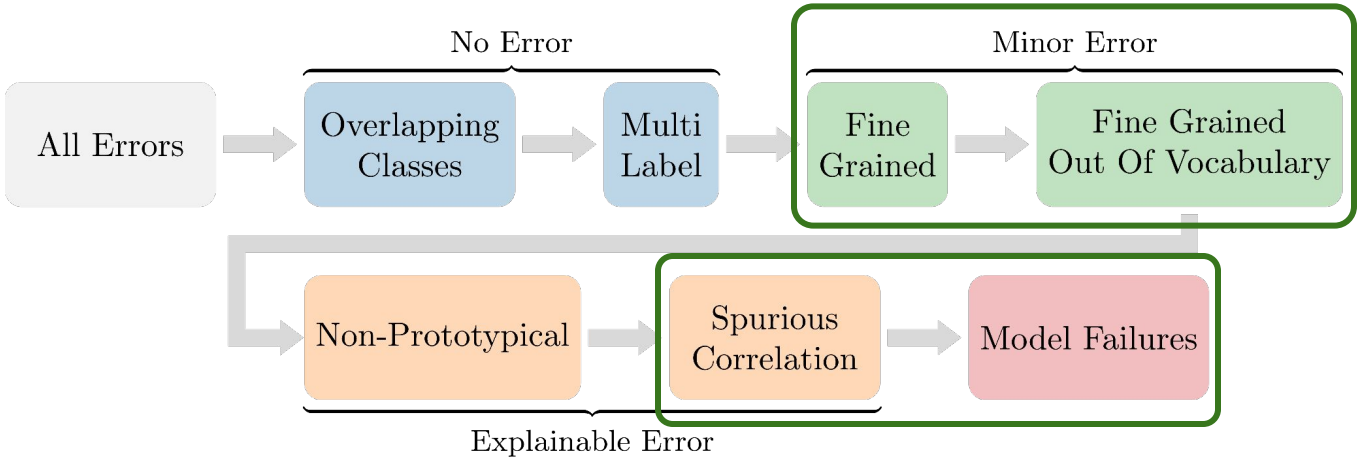
- ✓ all error categories identified by prior work
- ✓ minimal-severity bias
- ✓ consistent and repeatable

⇒ study the *error distributions* of 900+ models

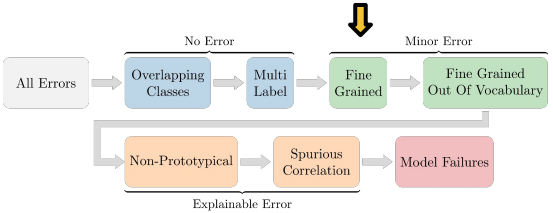
# **Automated** Classification of Model Errors



# **Automated** Classification of Model Errors

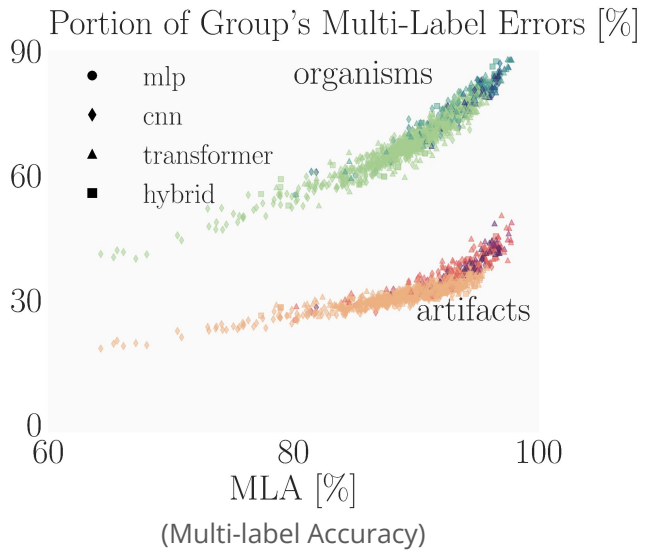


# Fine-Grained Errors



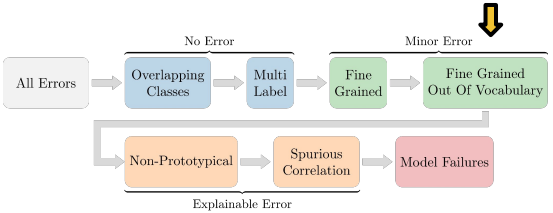

- Confuse similar, semantically related ImageNet classes
- Manually group all 1000 ImageNet classes into 161 superclasses

✓ Ground-truth: `tabby cat`
✗ Prediction: `Egyptian cat`
Same superclass: `domestic cat`



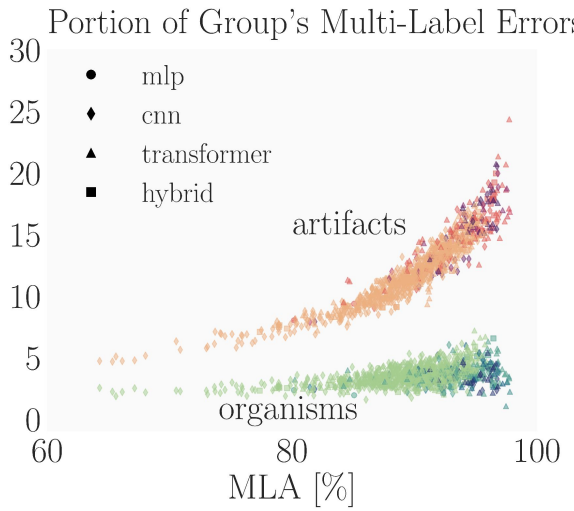

(Multi-label Accuracy)

# Fine-Grained OOV Errors

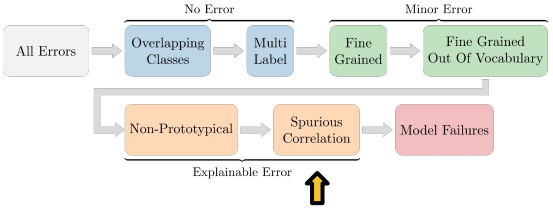



- Classify a prominent entity not in the ImageNet labelset
- Visually similar train sample in the same superclass → possibly a fine-grained error
- Collect proposals from WordNet and confirm OOV with an open world classifier


✓ Ground-truth: `coral reef`
✗ Prediction: `rock beauty`
OOV proposal: `butterflyfish`

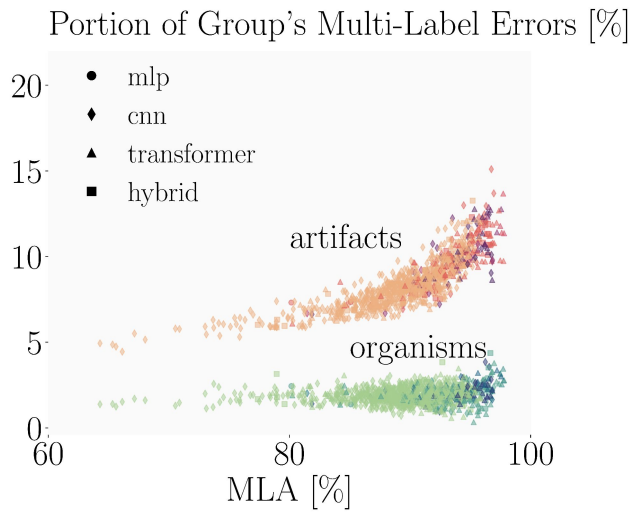


# Spurious Correlations

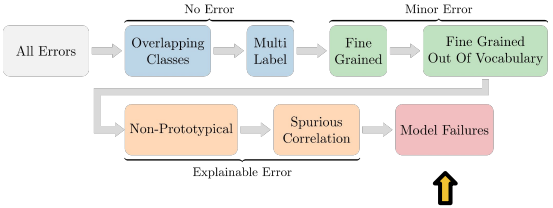



- Identify commonly co-occurring classes

✓ Multi-labels: `ski mask, alp`
✗ Prediction: `ski`

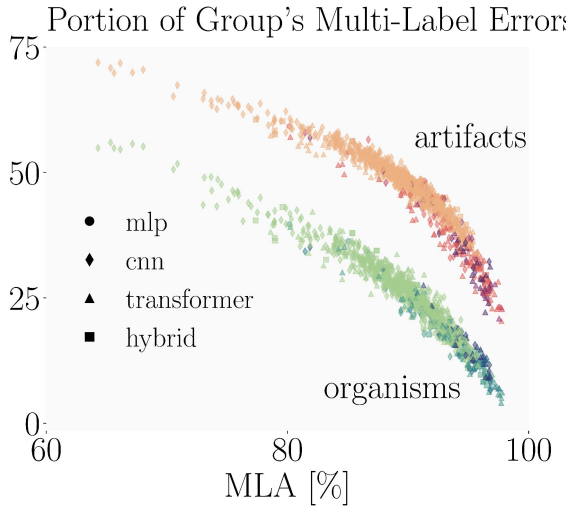


# Model Failures




- Particularly severe, hard to explain errors

✓ Multi-labels: `basket, hamper`
✗ Prediction: `pillow`



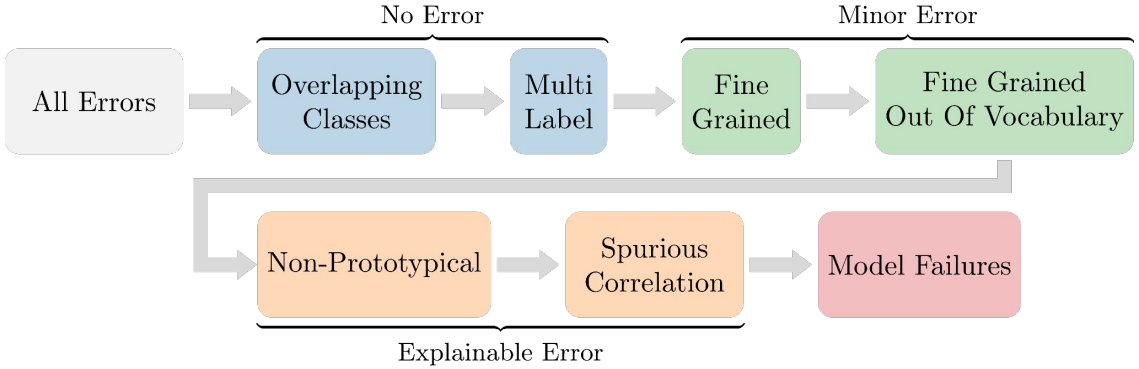

⇒ MLA pessimistic: model failures decrease faster than multi-label errors

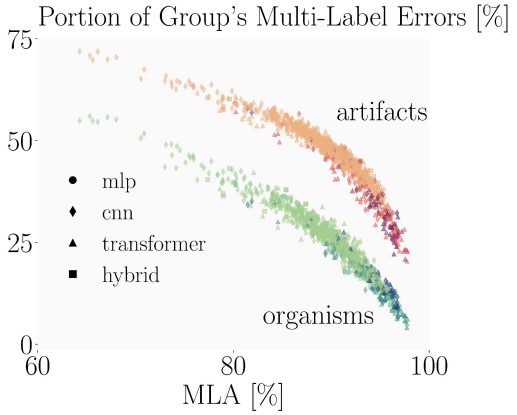
⇒ Portion of model failures higher for artifacts, but drops rapidly

# Further details in the paper:

- Model pre-training datasets
- Model architecture
- Alignment to human experts
- Extension to other datasets

# Summary




Code, evaluation & analysis:

https://github.com/eth-sri/automated-error-analysis