Imperfect and uncertain labels in computer vision

Gala

gegeh@create.aau.dk

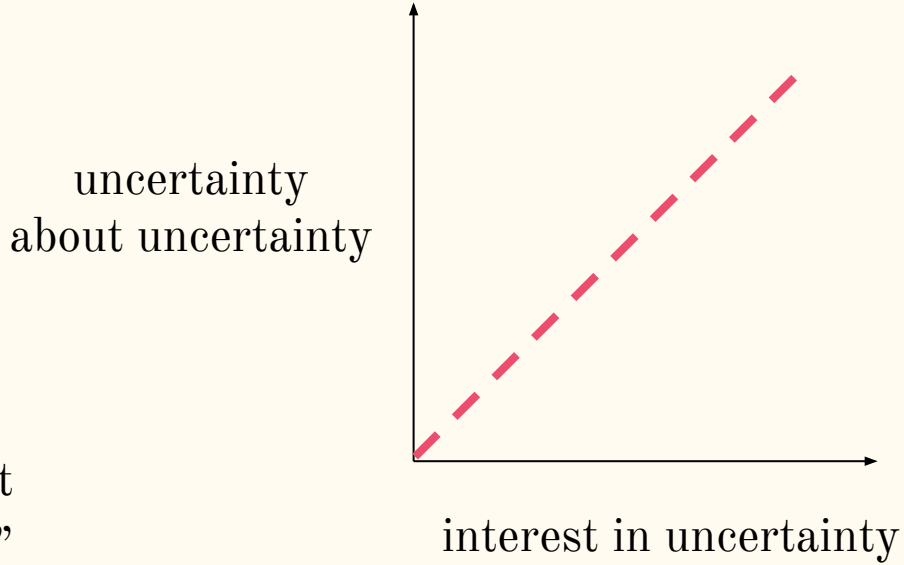
AALBORG UNIVERSITY

VISUAL ANALYSIS & PERCEPTION LAB

About me

PhD “imperfect labels and imperfect classifiers”

Now: chaotic postdoc, CV + NLP



Agenda

- what is aleatoric uncertainty
- labels in computer vision, and why focus on that
- sources of label ambiguity, uncertainty and errors
- how it manifests in practice across different computer vision tasks and domains
- how it can be measured and modelled
- how imperfect training labels can affect model performance and post-hoc tasks such as OOD detection
- the importance of also taking into account imperfections or uncertainty in the labels used for evaluation

What is aleatoric uncertainty

Aleatoric uncertainty is...

- When there are multiple plausible labels y for a given input x:

  entropy of $\mathbf{P(Y|X=x) > 0}$ (true data generating distribution)

- A property of the data
- Independent of the model
- What remains when the epistemic uncertainty goes to 0 (perfect model)
- Often called “irreducible” uncertainty

1. **Epistemic uncertainty**: “I am not sure because I have not seen it before.”
2. **Aleatoric uncertainty**: “I have experienced it before, I know what I am doing, but I think there is more than one good answer to your question, so I cannot choose just one.”

Aleatoric uncertainty is...

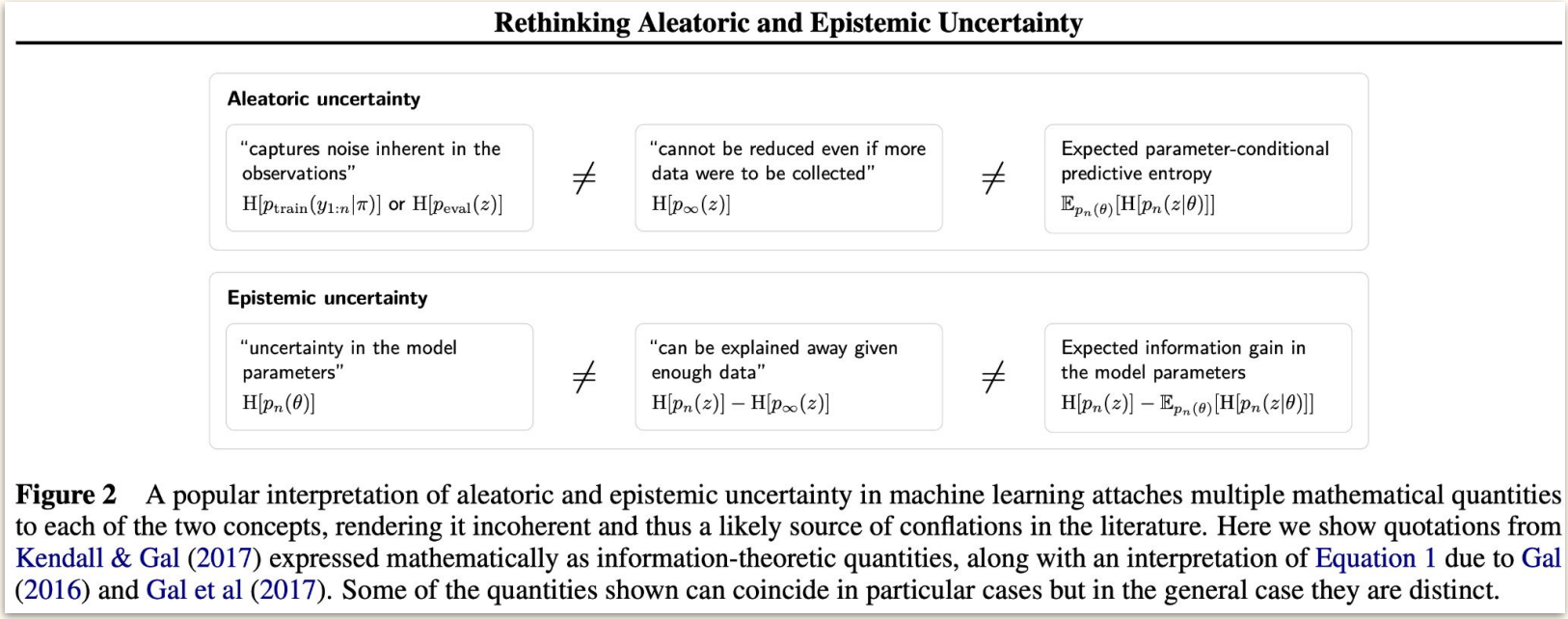


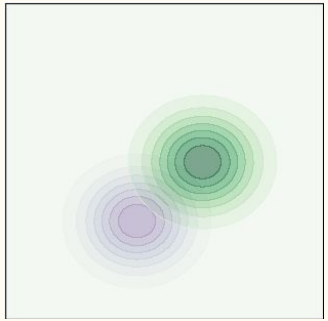
Figure 2 A popular interpretation of aleatoric and epistemic uncertainty in machine learning attaches multiple mathematical quantities to each of the two concepts, rendering it incoherent and thus a likely source of conflations in the literature. Here we show quotations from Kendall & Gal (2017) expressed mathematically as information-theoretic quantities, along with an interpretation of Equation 1 due to Gal (2016) and Gal et al (2017). Some of the quantities shown can coincide in particular cases but in the general case they are distinct.

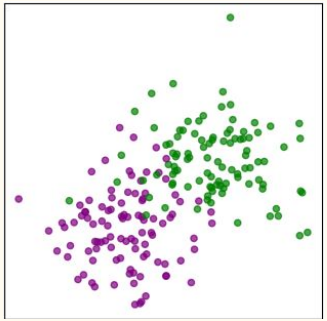
Rethinking Aleatoric and Epistemic Uncertainty ICML, 2025
Reexamining the Aleatoric and Epistemic Uncertainty Dichotomy | ICLR Blogposts 2025

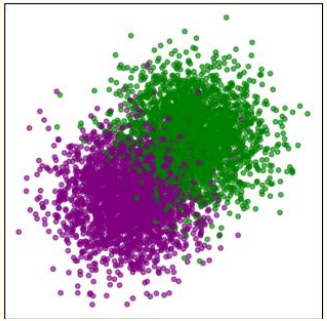
Can aleatoric uncertainty be reduced?

Aleatoric is often called **“irreducible” uncertainty**

Is indeed **irreducible if the data/label generation process stays fixed**: observing more samples does not help



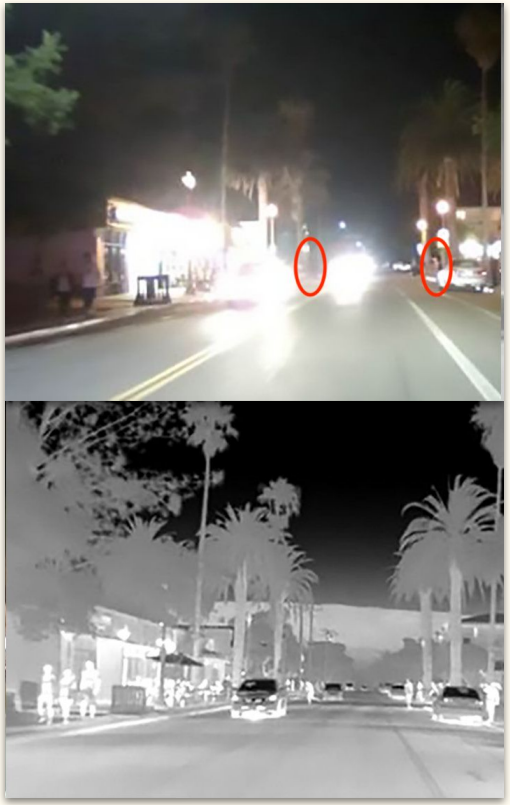




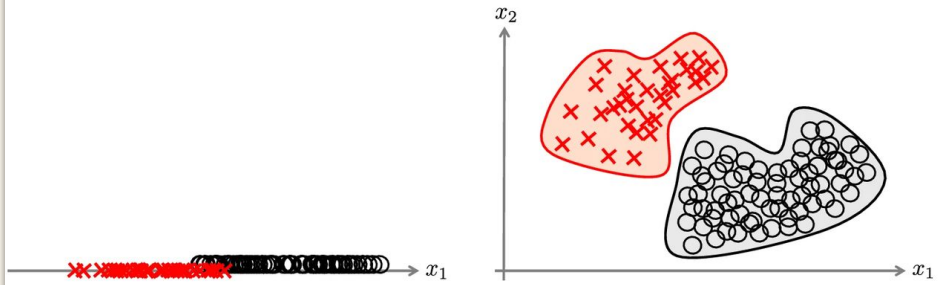
Can aleatoric uncertainty be reduced?

Aleatoric is often called **“irreducible” uncertainty**

However, it can be reduced by **changing the data generation process**: e.g. adding (informative) features, another modality, using a less noisy sensor, hiring more meticulous annotators etc.



From: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods



Left: The two classes are overlapping, which causes (aleatoric) uncertainty in a certain region of the instance space. Right: By adding a second feature, and hence embedding the data in a higher-dimensional space, the two classes become separable, and the uncertainty can be resolved

Wisdom from NLP

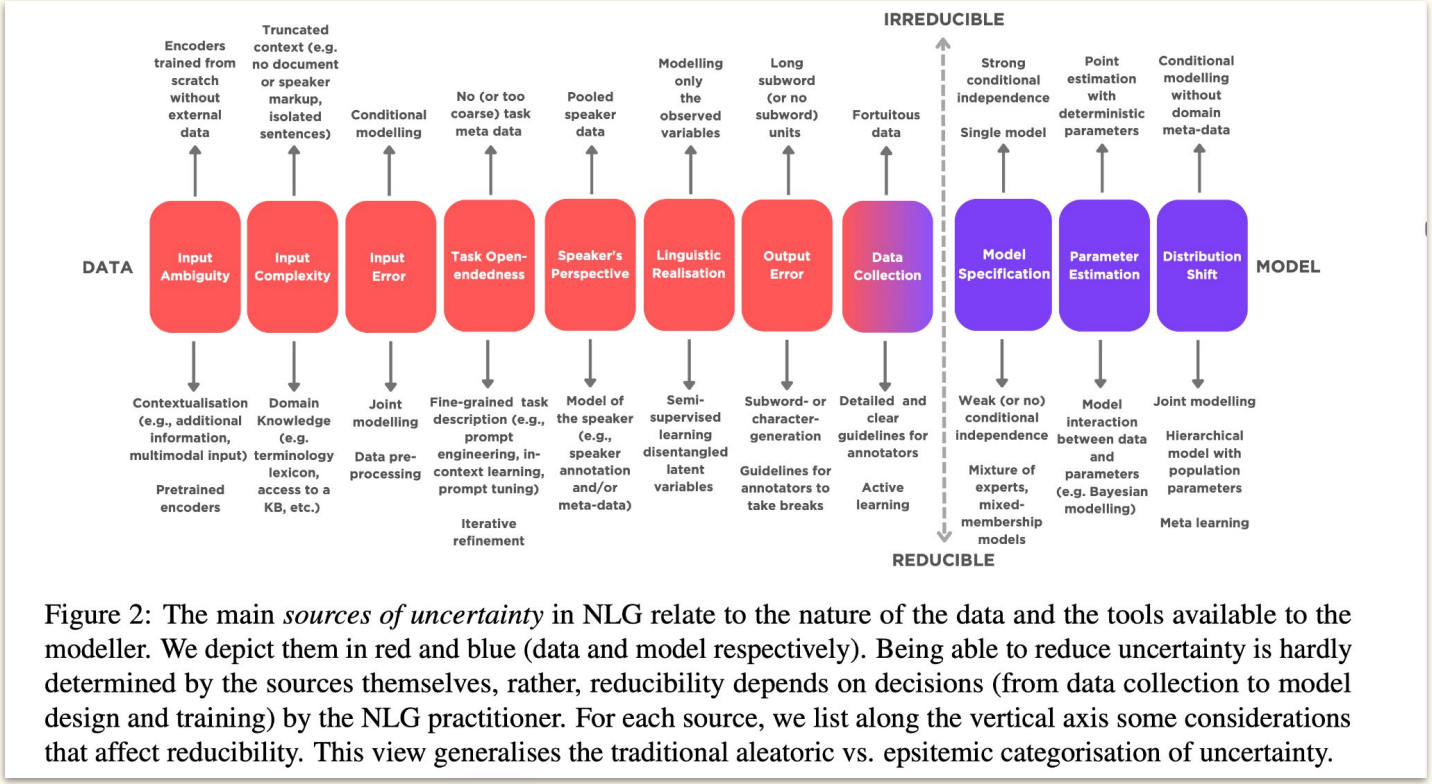


Figure 2: The main *sources of uncertainty* in NLG relate to the nature of the data and the tools available to the modeller. We depict them in red and blue (data and model respectively). Being able to reduce uncertainty is hardly determined by the sources themselves, rather, reducibility depends on decisions (from data collection to model design and training) by the NLG practitioner. For each source, we list along the vertical axis some considerations that affect reducibility. This view generalises the traditional aleatoric vs. epsitemic categorisation of uncertainty.

[2307.15703] Uncertainty in Natural Language Generation: From Theory to Applications

Labels in computer vision

Why focus on data/labels?

How often have you used a benchmark dataset without browsing individually labeled samples?

How often do dataset papers describe the annotation protocol in detail or potential sources of ambiguity/error?

If you've annotated a dataset yourself, how often did you hesitate during the process?

In computer vision, especially "general" CV, we like to **pretend that labels are certain and reliable**, both in training and in testing.

Reality: the labeling process is imperfect and does not necessarily recover the "true" label. It is only an approximation of the true data generation process

Why focus on data/labels?

Data is the bread and butter of deep learning, both for **training** and **evaluation**

Characterizing errors or ambiguity in the data can help better understand the task/problem, develop better models and evaluate them more fairly.

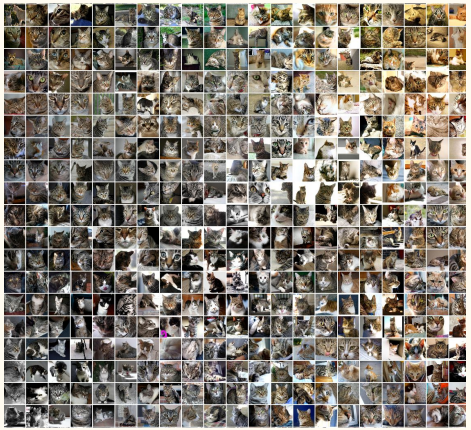
Before training	**During** training	**After** training
Characterize noise / uncertainty	Down-weigh ambiguous or mislabeled samples	Nuance the evaluation based on mistake severity
Use these insights to inform the choice of model / training objective	Leverage annotation metadata / privileged information	Estimate the aleatoric uncertainty during deployment

Scope of this talk

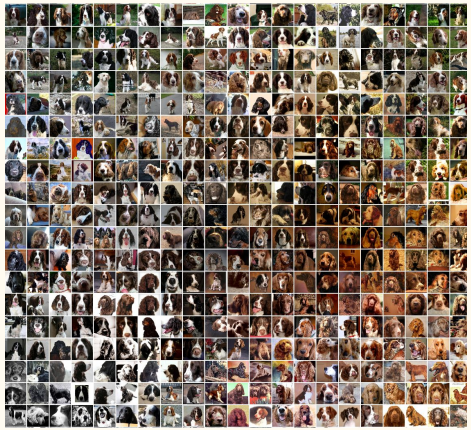
Main focus: multi-class supervised learning, with image-level or pixel-level labels.

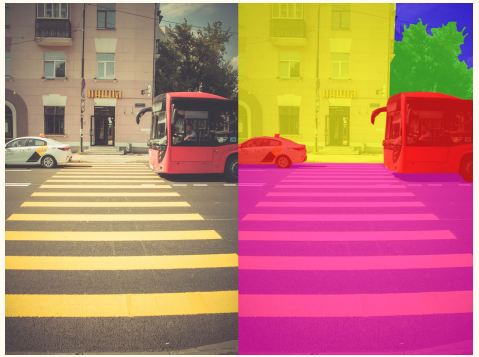
But many of the points/observations apply to other CV tasks.

label: cat



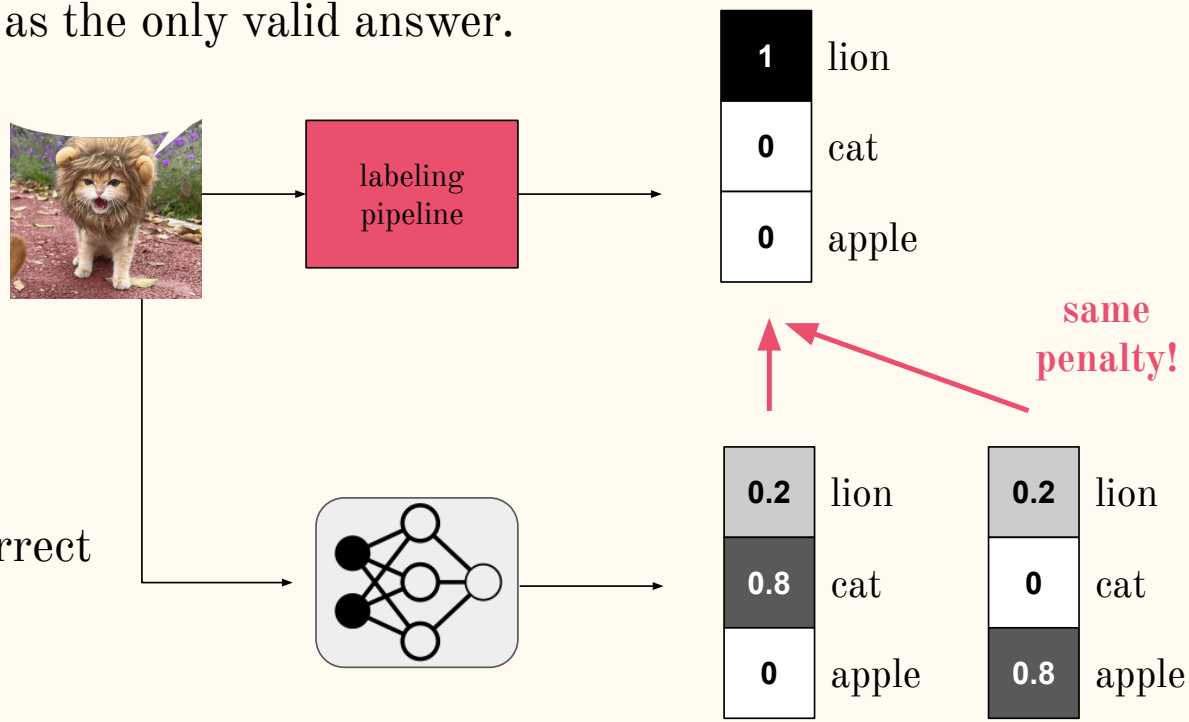
label: dog





Spectrum of label imperfection/ambiguity

The typical training and evaluation pipeline treats every label as 100% correct, and as the only valid answer.



In reality:

- the assigned label may be incorrect
- other labels may be correct or at least acceptable

Spectrum of label imperfection/ambiguity

100%, unjustifiably/unacceptably wrong

wrong but difficult/not surprising

one of several valid interpretations

100% correct, everyone would agree

Interactive catalog of ImageNet labeling errors: https://labelerrors.com/

Flaws of ImageNet, Computer Vision's Favorite Dataset | ICLR Blogposts 2025

Sources of label ambiguity and label errors

The correctness and certainty of a labeled image depends on many factors

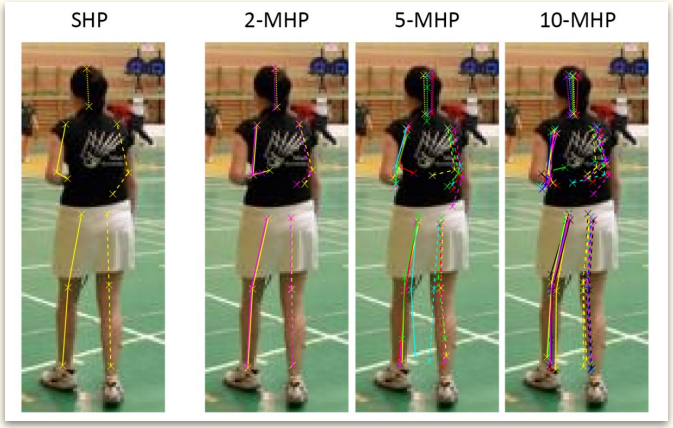
- whether the image contains sufficient (discriminative) information
- the level of expertise/knowledge/effort of human labellers
- the task and class definition itself
- the design of the annotation task/tool/interface
- whether and how certain parts of the annotation process have been automated

[2205.04596] When does dough become a bagel? Analyzing the remaining mistakes on ImageNet 2022

Insufficient information in the image

- Cropping or occlusion by another object.
- Blurry, noisy, low-contrast, low-resolution images.
- Adverse environmental conditions (weather, turbidity etc)

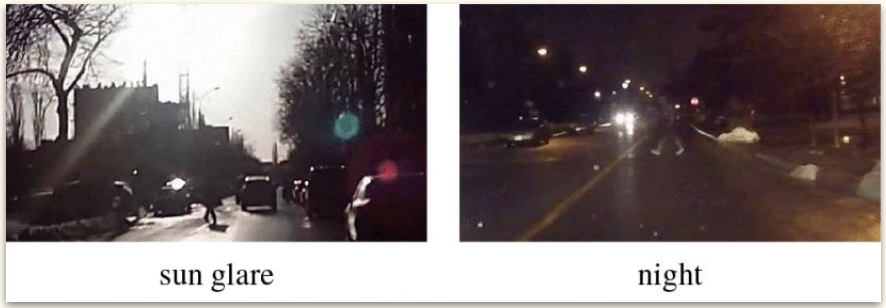
ImageNet1K image
monastery? church? palace? triumphal arch?



[1612.00197] Learning in an Uncertain World: Representing Ambiguity Through Multiple Hypotheses

Insufficient information in the image

- Cropping or occlusion by another object.
- Blurry, noisy, low-contrast, low-resolution images.
- Adverse environmental conditions (weather, turbidity etc)



Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior ICCVW 2017

The Brackish Dataset | vap.aau.dk ICCVW, 2019

Knowing and caring enough

- Insufficient **knowledge/experience** or **effort/incentive** can be a bottleneck in the quality and certainty of manual annotations.
- Malicious or accidental **mistakes (e.g. mis-clicking)** can also occur.
- Reducing the time/money spent per label allows for larger datasets to be labeled, but comes with a **label quality trade-off**.

[2303.15850] That Label's Got Style: Handling Label Style Bias for Uncertain Image Segmentation ICLR, 2023

Mismatch between task and data

The labeling task/class definition imposes a **set of assumptions on the data which might not be met in practice**. This will introduce incorrect or partially incorrect labels no matter how clear the images is & how dedicated the annotator is.

e.g. "pick the correct label: dog or cat"

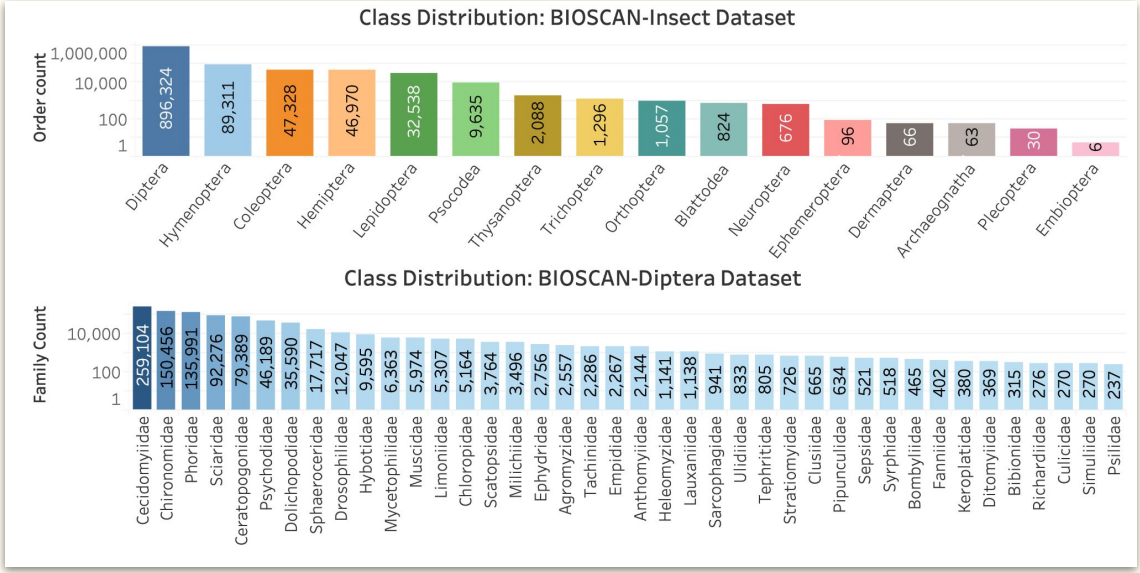
multi-label images, even though task imposes a single label per image

none of the pre-defined categories in the image

Class definition

The set of classes of interest heavily influences the inherent difficulty, ambiguity, or subjectivity of the task.

- ➔ more fine-grained: more difficult to distinguish
- ➔ ordinal ranks
- ➔ subjective ratings
- ➔ fuzzy boundaries between classes, arbitrary discretization



[2307.10455] A Step Towards Worldwide Biodiversity Assessment: The BIOSCAN-1M Insect Dataset NeurIPS 2023

Class definition

The set of classes of interest heavily influences the inherent difficulty, ambiguity, or subjectivity of the task.

- ➔ more fine-grained: more difficult to distinguish
- ➔ ordinal ranks — e.g. age estimation, tumor grading
- ➔ subjective ratings — e.g. image quality assessment, hatefulness rating
- ➔ fuzzy boundaries between classes, arbitrary discretization — e.g. sidewalk vs. road, fashion styles, art genres, foreground vs. background

Class definition

Examples of potentially overlapping class pairs in ImageNet

sunglass	sunglasses, dark glasses, shades
projectile, missile	missile
dough	bagel
sofa	studio couch
convertible	sports car

Label: dough; Model: bagel. When does dough become a bagel?

IMAGENET 1000 Class List
Browse ImageNet interactively: NAVIGU
[2205.04596] When does dough become a bagel? Analyzing the remaining mistakes on ImageNet NeurIPS 2022

Boundaries in time and space

Most uncertainty in segmentation labels occurs along boundaries. Even when the image is clear, there is a certain fuzziness in delineating where an action, object or part starts and ends

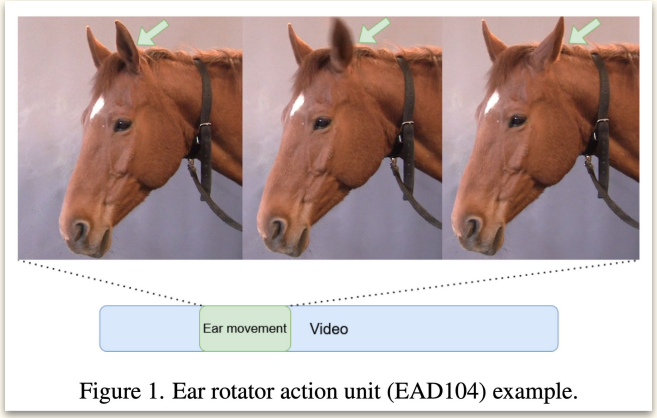
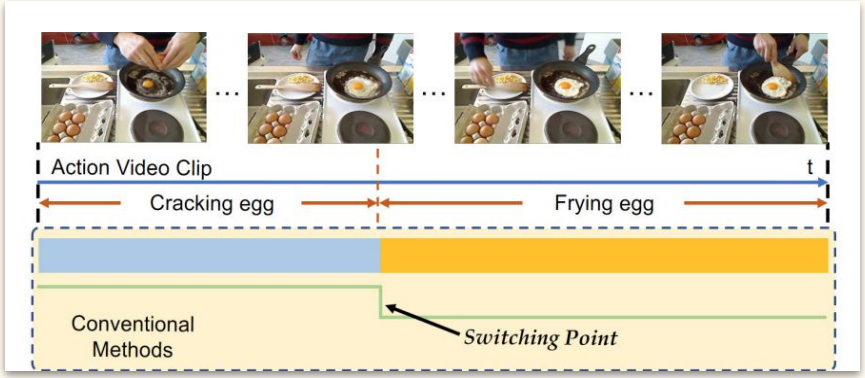


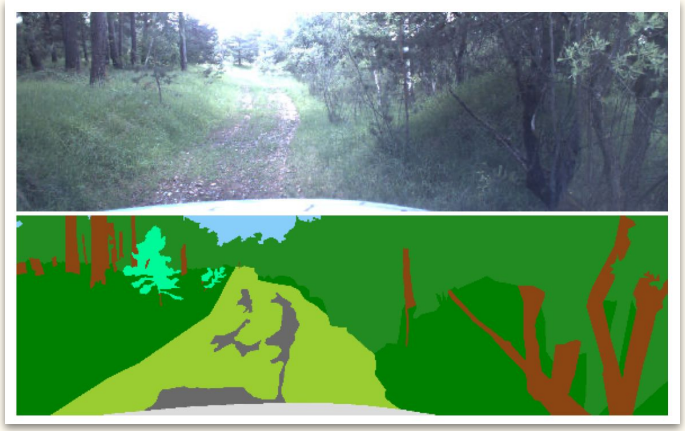
Figure 1. Ear rotator action unit (EAD104) example.

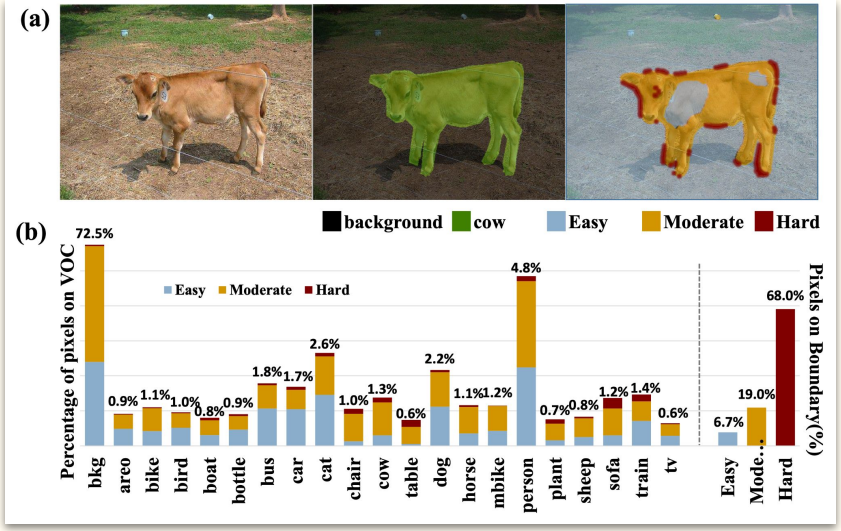


[2505.03554] Read My Ears! Horse Ear Movement Detection for Equine Affective State Assessment CVPRW 2025
Uncertainty-Aware Representation Learning for Action Segmentation | IJCAI

Boundaries in time and space

Most uncertainty in segmentation labels occurs along boundaries. Even when the image is clear, there is a certain fuzziness in delineating where an action, object or part starts and ends





[2103.13109] A Fine-Grained Dataset and its Efficient Semantic Segmentation for Unstructured Driving Scenarios ICPR, 2021
[1704.01344] Not All Pixels Are Equal: Difficulty-aware Semantic Segmentation via Deep Layer Cascade CVPR, 2017

Design choices influence label bias & variance

task & interface

- how is the task formulated?
- what instructions are given? how specific are they?
- what exemplars/edge cases are shown? (e.g. how to deal with occlusions, reflections, tiny objects)
- how is the UI designed?
- can the image be adjusted?

participants

- how tedious/long is the task? is it split into several sessions?
- what device(s) is used?
- what is the incentive structure?

Design choices influence label bias & variance

Browsing vs. tagging interface for image classification

browsing interface presents a **single concept along with a set of candidate images** arranged in a grid and asks the annotator to select the images correctly depicting the concept.

e.g. **ImageNet**, Places, and CUB.

tagging interface presents a **single image at a time** and asks the annotator to choose one or more objects and concept labels as necessary.

e.g. Pascal, COCO, LVIS, and iNaturalist.

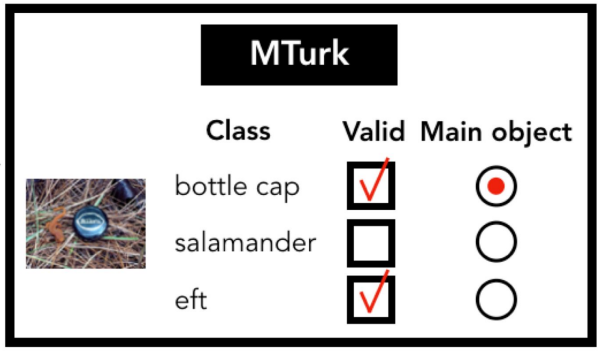
[2104.08885] A survey of image labelling for computer vision applications Journal of Business Analytics, 2021.
[2303.17595] Neglected Free Lunch -- Learning Image Classifiers Using Annotation Byproducts ICCV, 2023.

Design choices influence label bias & variance

when asked to select images containing a specific class, ImageNet annotators give significantly different answers than when asked to select the correct class for a given image

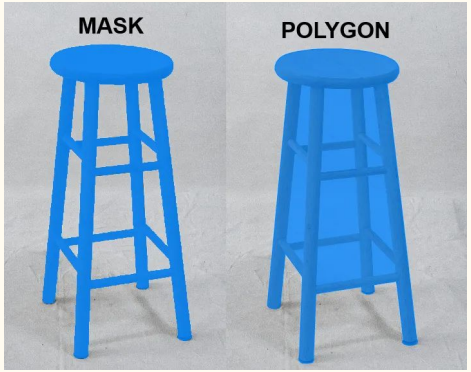
[2005.11295] From ImageNet to Image Classification: Contextualizing Progress on Benchmarks ICML, 2020

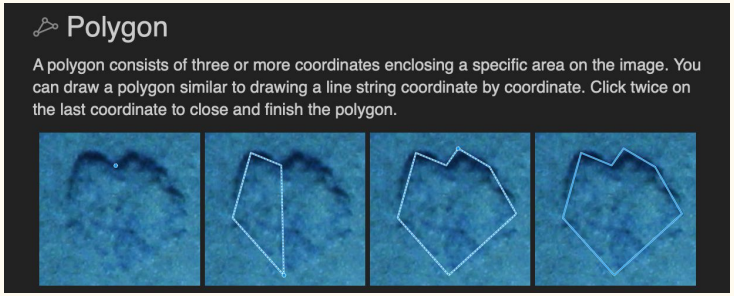
vs.

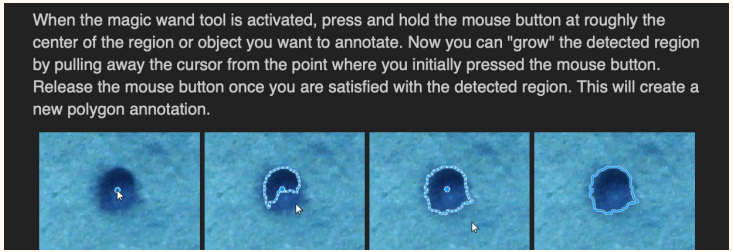


Design choices will influence label bias & variance

e.g. different segmentation annotation tools will lead to different annotations







Creating Image Annotations - BIIGLE
Brush (Mask) Tool in CVAT for Segmentation | CVAT Academy

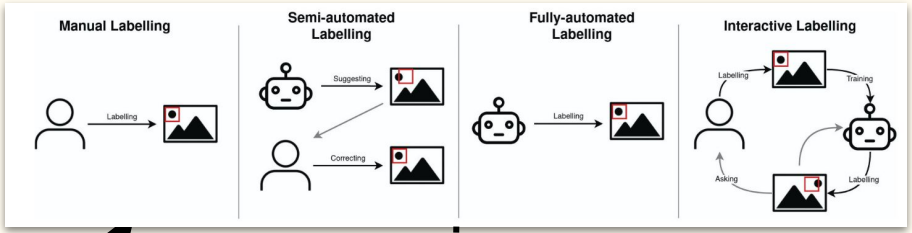
Automation in the annotation pipeline

raw images / data sources

which samples should be labeled?

curated collection

how are they labeled?



data selection and filtering

e.g. auto label propagation or label proposals can introduce additional errors

[2104.08885] A survey of image labelling for computer vision applications Journal of Business Analytics, 2021.

Wisdom from NLP

"**Errors** can already be present in the **data sources** themselves"
"Moreover, the **automatic pre-processing** steps are prone to errors themselves, which are often not transparent to annotators and user"
"Manual annotation is time-consuming and cognitively heavy. **Human errors** might therefore occur and not be detected, even in iterative rounds"

"**Uncertainty** arises wherever multiple possible interpretations of data present themselves, but the **relevant knowledge or information to unequivocally opt for one of the interpretations is not available**"

"**Ambiguity** occurs whenever an **entity in principle allows for multiple interpretations.**"

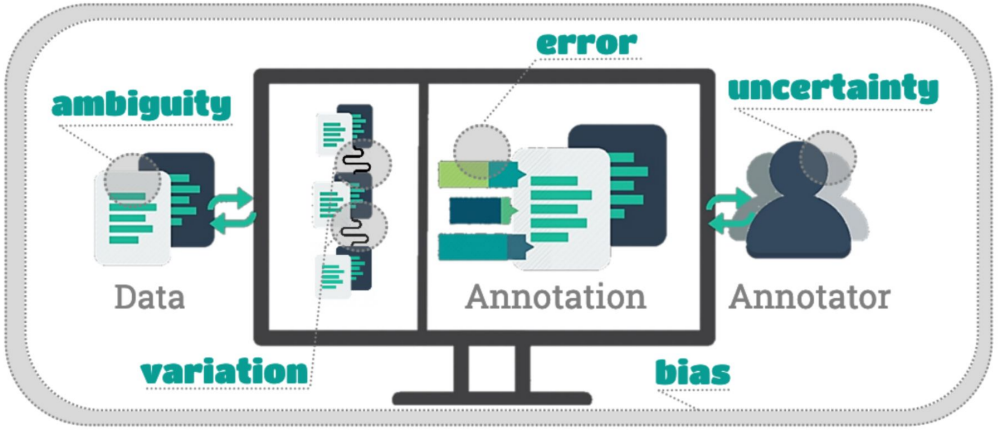


Figure 1: Representation problems in linguistic annotations come from five distinct sources: (i) **Ambiguities** are an inherent property of the *data*. (ii) **Variation** is also part of the *data* and can, e.g., occur across documents. (iii) **Uncertainty** is introduced by an *annotator*'s lack of knowledge or information. (iv) **Errors** can be found in the *annotations*. (v) **Biases** are a property of the *complete annotation system*.

"**Variation** is not an annotation issue in terms of identification and classification,
but rather that **a single interpretation (variable) manifests itself in two or more (variants).**"

Bias represents an **influence which leads to a preference or tendency for one thing** over another. Often, neither the end-user nor the annotator may be conscious of such biases

Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias - ACL Anthology

Wisdom from NLP

"Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation" The AI Magazine, 2015

Myth One: One Truth

Most data collection efforts assume that there is one correct interpretation for every input example. The **ideal of truth is a fallacy for semantic interpretation** and needs to be changed.

Myth Two: Disagreement Is Bad

Disagreement has been **considered a measure of poor quality** in the annotation task, either because the task is poorly defined or because the annotators lack sufficient training. However, in our research we found that **disagreement is not noise but signal**, and at the level of individual examples can indicate that the sentence or the relation at hand is ambiguous or vague, or that the worker is not doing a good job

Myth Three: Detailed Guidelines Help

Increasingly precise annotation guidelines **do eliminate disagreement** but **do not increase quality**, perfuming the agreement scores by forcing human annotators to make choices they may not actually think are valid, and removing the potential signal on individual examples that are vague or ambiguous.

Wisdom from NLP

"Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation" The AI Magazine, 2015

Myth Four: One Is Enough

Due to the time and cost required to generate human annotated data, standard practice is for the vast majority, often more than 90 percent, of annotated examples to be seen by a single annotator. We see many examples where just **one perspective isn't enough**.

Myth Five: Experts Are Better

Conventional wisdom is that if you want medical texts annotated for medical relations you need medical experts. In our work, experts did not show significantly better quality annotations than non-experts. Crowd annotators in general tended to read the sentences more literally, which made them better suited to provide examples to a machine. We have also found that **multiple perspectives on data, beyond what experts may believe is salient or correct, can be useful**.

Wisdom from NLP

"Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation" The AI Magazine, 2015

Myth Six: All Examples Are Created Equal

The mathematics of using ground truth treats every example the same; either you match the correct result or not. We can find by recording disagreement on each example that poor quality examples tend to generate high disagreement. The disagreement allows us to weight some samples higher than others, giving us the ability to both **train and evaluate a machine in a more flexible way**.

Myth Seven: Once Done, Forever Valid

Perspectives change over time, which means that training data created years ago might contain examples that are not valid or only partially valid at a later point in time.

"Crowd truth has allowed us to identify and dispel the myths of human annotation and **to paint a more accurate picture of human performance on semantic interpretation for machines to attain."**

Characterizing uncertainty at the data level

Characterizing uncertainty at the data level

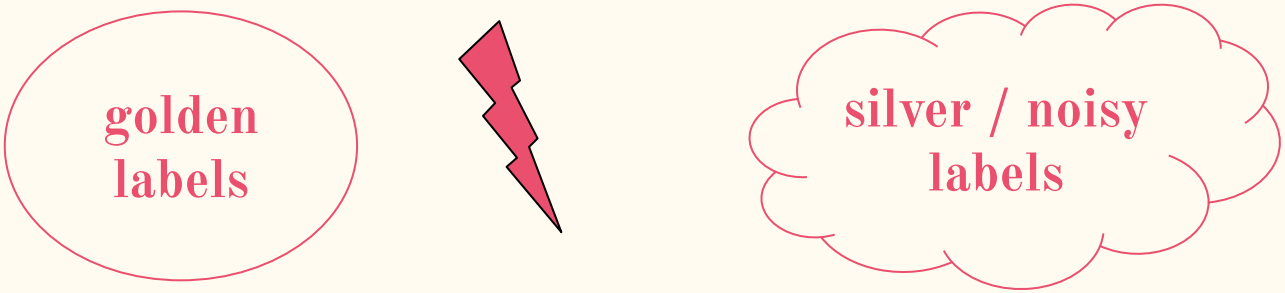
Let's say you're collecting a dataset. What are different ways to characterize the quality/difficulty/ambiguity of the data?

- **noisy labels vs. gold standard:** each sample is associated with one silver label, but you also have access to some kind of golden reference for some/all samples
- **multi-annotator labels:** each sample is annotated more than once (e.g. different people, same person at different times)
- **self-reported confidence:** each sample is associated with a label + confidence for that label
- **other metadata/byproducts** e.g. annotation time, behaviour, which can serve as a proxy for confidence/difficulty

Characterizing label noise

Let's say you're collecting a dataset. What are different ways to characterize the quality/difficulty/ambiguity of the data?

- **noisy labels vs. gold standard:** each sample is associated with one silver label, but you also have access to some kind of golden reference for some/all samples



costly or intrusive
(e.g. biopsy, domain expert)

cheap and scalable
(e.g. crowd-sourcing, auto-labeling)

Characterizing label noise

Let's say you're collecting a dataset. What are different ways to characterize the quality/difficulty/ambiguity of the data?

- **noisy labels vs. gold standard:** each sample is associated with one silver label, but you also have access to some kind of golden reference for some/all samples

> "Based on a random sample of 504 images, a board-certified dermatologist identified 69.0% of images as diagnostic of the labeled condition, 19.2% of images as potentially diagnostic (not clearly diagnostic but not necessarily mislabeled, further testing would be required), 6.3% as characteristic (resembling the appearance of such a condition but not clearly diagnostic), 3.4% are considered wrongly labeled, and 2.0% are labeled as other. A second board-certified dermatologist also examined this sample of images and confirmed the error rate." Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset CVPRW 2021

Characterizing label noise

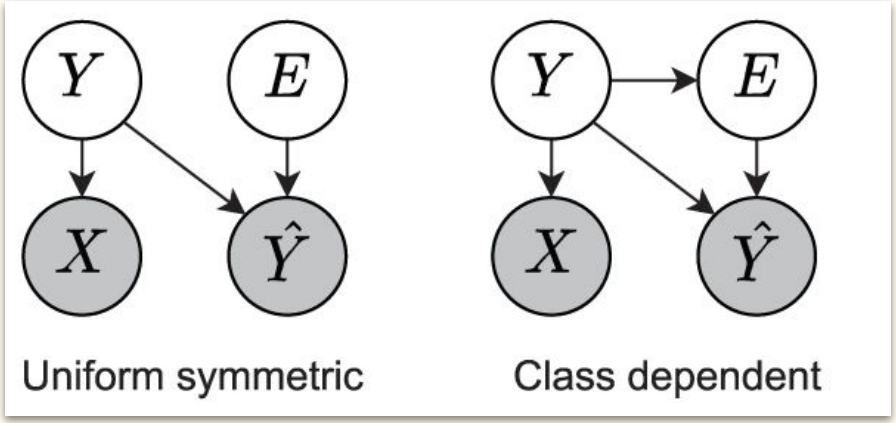
Properties of noisy labels:

Noise rate: This is the proportion of samples in a dataset which are mislabeled. A noise rate of 0% indicates clean labels.

Noise model: This describes the relationship between the occurrence of a labeling error and different attributes of the data.

Closed-set vs. open-set label noise: The open-set noisy label setting is more general: it considers that samples of unknown classes can also have inadvertently ended up in the dataset (these are mislabeled by definition)

Characterizing label noise: **noise model**



input image (X)
true label (Y)
observed label (Ŷ)
error occurrence (E)

noisy completely at random (NCAR)

all samples are **equally likely** to be mislabeled

noisy at random (NAR)

some **classes** are more likely to be mislabeled

Classification in the Presence of Label Noise: A Survey TNNLS 2013

Active label cleaning for improved dataset quality under resource constraints | Nature Communications 2022

Characterising label noise: **noise model**

The noise transition matrix summarizes random mislabeling patterns

Uniform/symmetric/NCAR

But labeled as...

True label was

	cat	dog	car	dock	rock
cat	0.8	0.05	0.05	0.05	0.05
dog	0.05	0.8	0.05	0.05	0.05
car	0.05	0.05	0.8	0.05	0.05
dock	0.05	0.05	0.05	0.8	0.05
rock	0.05	0.05	0.05	0.05	0.8

Class-dependent/asymmetric/NAR

But labeled as...

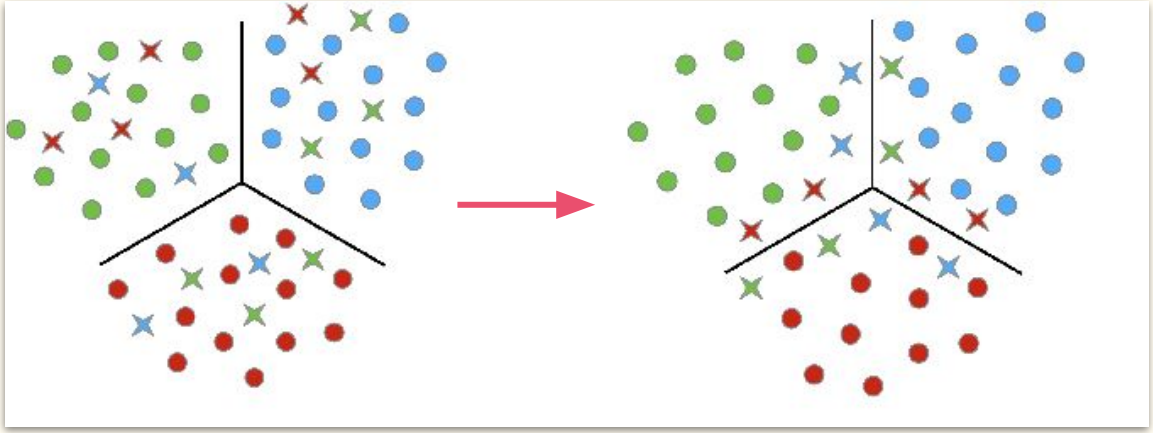
True label was

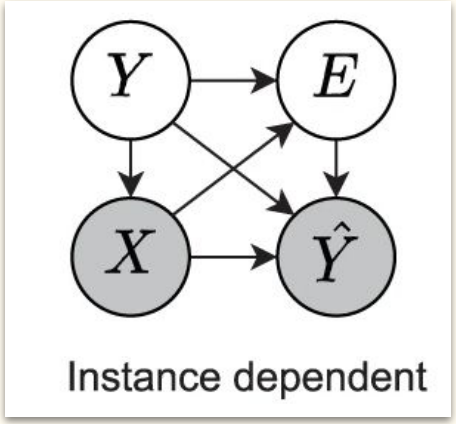
	cat	dog	car	dock	rock
cat	0.7	0.1	0.1	0.05	0.05
dog	0	0.9	0	0.1	0
car	0.1	0	0.8	0.05	0.05
dock	0	0	0	0.6	0.4
rock	0	0	0	0	1

Characterizing label noise: **noise model**

In practice, “real-world” label noise follows a more complex pattern: some **instances** are more likely to be mislabeled.

input image (X)
true label (Y)
observed label ($\hat{Y}$)
error occurrence (E)





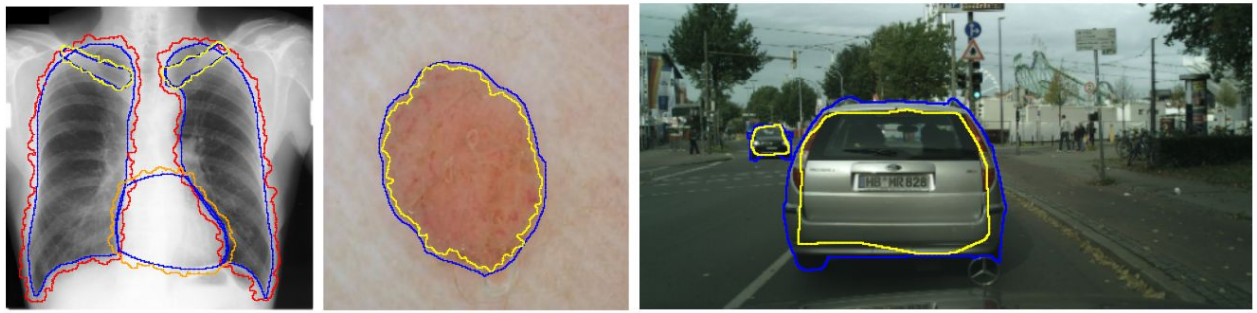
[2001.03772] Confidence Scores Make Instance-dependent Label-noise Learning Possible ICML, 2021

noisy not at random (NNAR)

Characterizing label noise: **noise model**

In **pixel-wise labeling tasks** (e.g. semantic segmentation), we don't label pixels independently. Label noise is highly likely to be **spatially correlated!**

Fig. 2. From left to right: unperturbed, natural perturbations, choppy perturbations, and random perturbations.



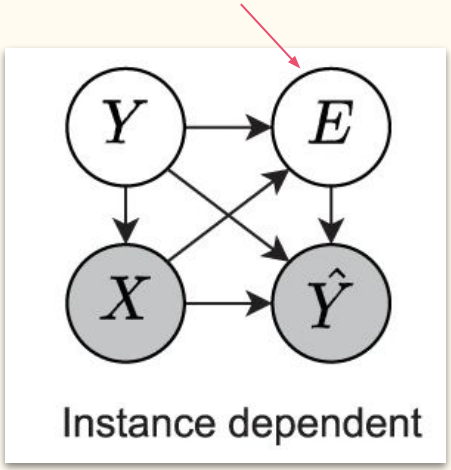
(a) *JSRT* S_E. (b) *ISIC 2017* S_S. (c) *Cityscapes*.

[1806.04618] Imperfect Segmentation Labels: How Much Do They Matter? LABELS, 2018
[2308.02498] Learning to Segment from Noisy Annotations: A Spatial Correction Approach ICLR, 2023

Characterizing label noise: **noise model**

In **pixel-wise labeling tasks** (e.g. semantic segmentation), we don't label pixels independently. Label noise is highly likely to be **spatially correlated!**

this also depends on the neighbouring pixels!



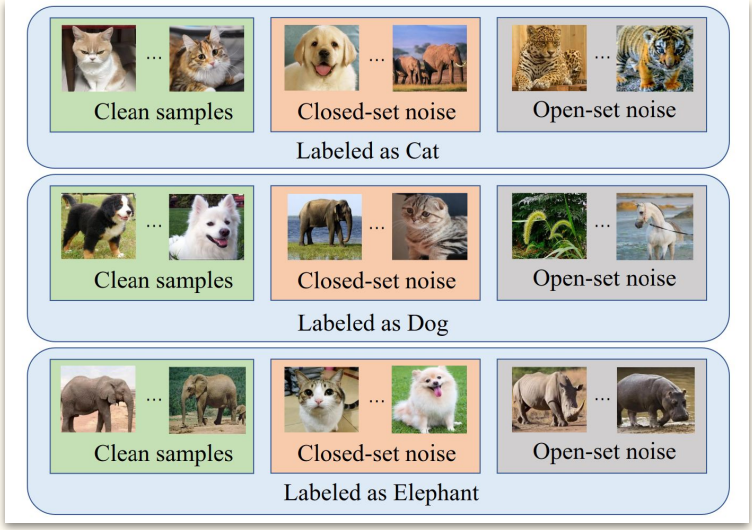
input pixel (X)
true label (Y)
observed label (^Y)
error occurrence (E)

Characterizing label noise: **open-set vs. closed-set**

Typical assumption: all curated samples belong to one of the pre-defined categories.

In reality, some data collection pipelines introduce unknown classes into the mix.

Unless there is an option to flag them or a catch-all “other” class, then they will inevitably mislabeled and introduce **open-set label noise.**



[2305.04203] Unlocking the Power of Open Set : A New Perspective for Open-Set Noisy Label Learning
AAAI, 2024.

Characterizing uncertainty at the data level

Let's say you're collecting a dataset. What are different ways to characterize the quality/difficulty/ambiguity of the data?

- **noisy labels vs. gold standard:** each sample is associated with one silver label, but you also have access to some kind of golden reference for some/all samples
- **multi-annotator labels:** each sample is annotated more than once (e.g. different people, same person at different times)
- **self-reported confidence:** each sample is associated with a label + confidence for that label
- **other metadata/byproducts** e.g. annotation time, behaviour, which can serve as a proxy for confidence/difficulty

Multi-annotator disagreement

Asking multiple people to annotate the same sample is standard practice in many fields (to measure **“inter-rater”/ “inter-observer” variability**), but not so much in “general” computer vision.

For a pair of annotators, the standard disagreement measure is Cohen’s kappa
$\boldsymbol{\kappa = (P_o - P_e) / (1 - P_e)}$

Where:

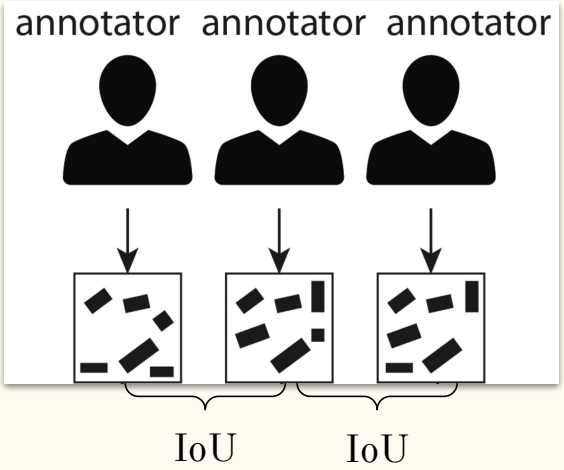
- $\mathbf{P_o}$ = observed agreement (proportion of samples both annotators agreed on)
- $\mathbf{P_e}$ = expected agreement by chance (based on each annotator's marginal label distribution)

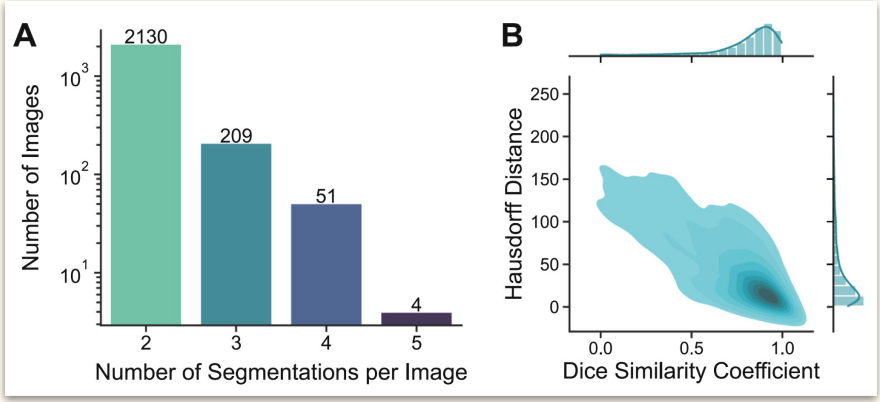
κ ranges from -1 to 1, where 1 = perfect agreement, 0 = agreement no better than chance, and negative values = worse than chance

There are other measures which generalize to >2 annotators (e.g. Fleiss’ **κ**)

Multi-annotator disagreement

It is also common to measure disagreement using **standard performance metrics** (e.g. accuracy, IoU, Dice), but **comparing pairs of annotators** rather than an annotator vs. the GT.

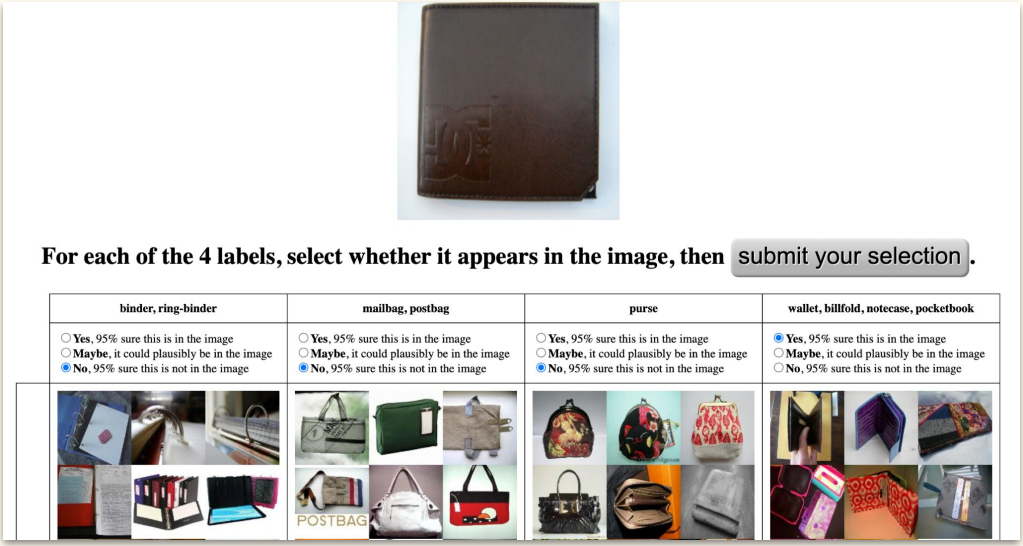




What Can We Learn from Inter-Annotator Variability in Skin Lesion Segmentation? DGM4MICCAI 2025

Self-reported confidence

You can give individual annotators the option to express their own uncertainty



[2006.07159] Are we done with ImageNet? 2020

Accurate integration of crowdsourced labels using workers' self-reported confidence scores IJCAI, 2013

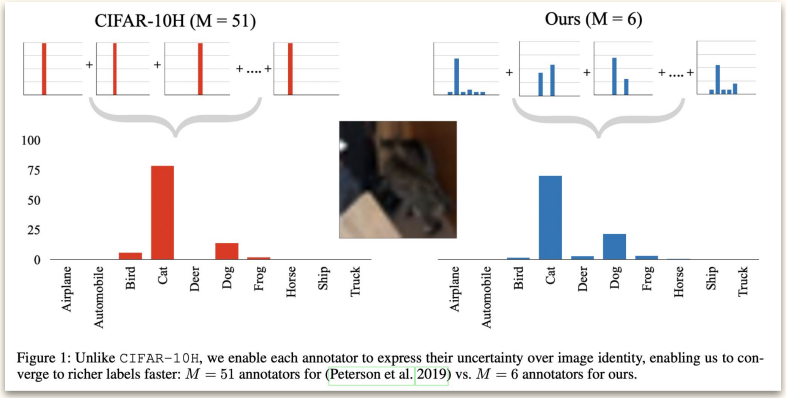


Figure 1: Unlike CIFAR-10H, we enable each annotator to express their uncertainty over image identity, enabling us to converge to richer labels faster: $M = 51$ annotators for (Peterson et al. 2019) vs. $M = 6$ annotators for ours.

[2207.00810] Eliciting and Learning with Soft Labels from Every Annotator AAAI, 2022

Annotation metadata

- time spent
- interaction with the application
- ID or expertise level

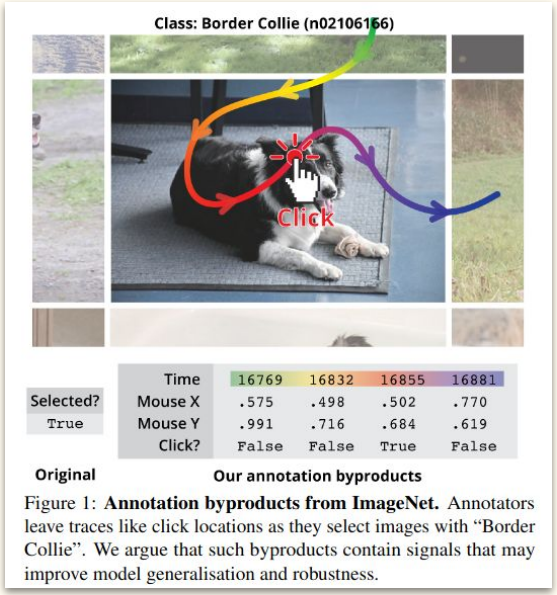


Figure 1: **Annotation byproducts from ImageNet.** Annotators leave traces like click locations as they select images with "Border Collie". We argue that such byproducts contain signals that may improve model generalisation and robustness.


```
"imageID": "n01440764/n01440764_105",
"originalImageHeight": 375,
"originalImageWidth": 500,
"selected": true,
"selectedRecord": [
  {"x": 0.540, "y": 0.473, "time": 1641425052}
],
"mouseTracking": [
  {"x": 0.003, "y": 0.629, "time": 1641425051},
  {"x": 0.441, "y": 0.600, "time": 1641425052}
]
```


Original Annotation

Annotation Byproducts

Figure 4: **Annotation byproducts from ImageNet.** See Appendix Figure A for the full list of byproducts.

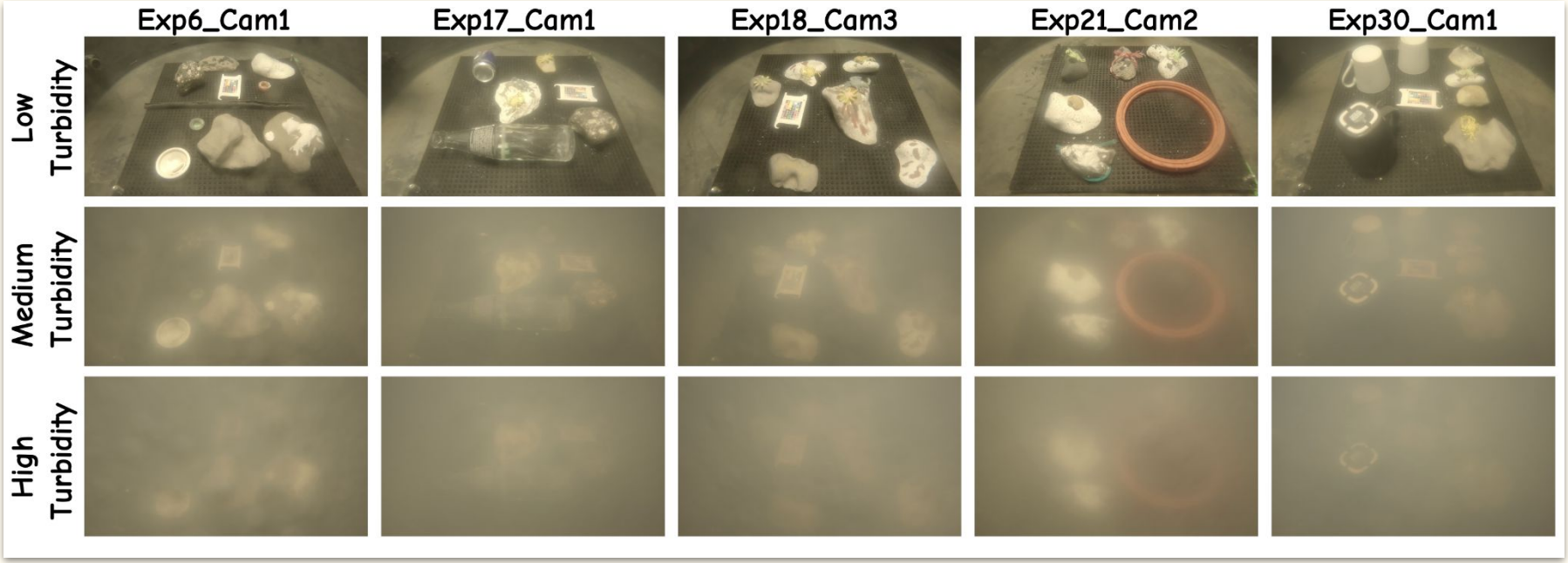
$$a = \begin{array}{r} \texttt{annotation time: 5 mins} \\ \texttt{confidence: 3\%} \\ \texttt{annotator id: \#00342} \\ \texttt{image id: \#1748} \end{array}$$

[2303.17595] Neglected Free Lunch -- Learning Image Classifiers Using Annotation Byproducts ICCV, 2023

[2310.06600] Pi-DUAL: Using Privileged Information to Distinguish Clean from Noisy Labels ICML, 2024

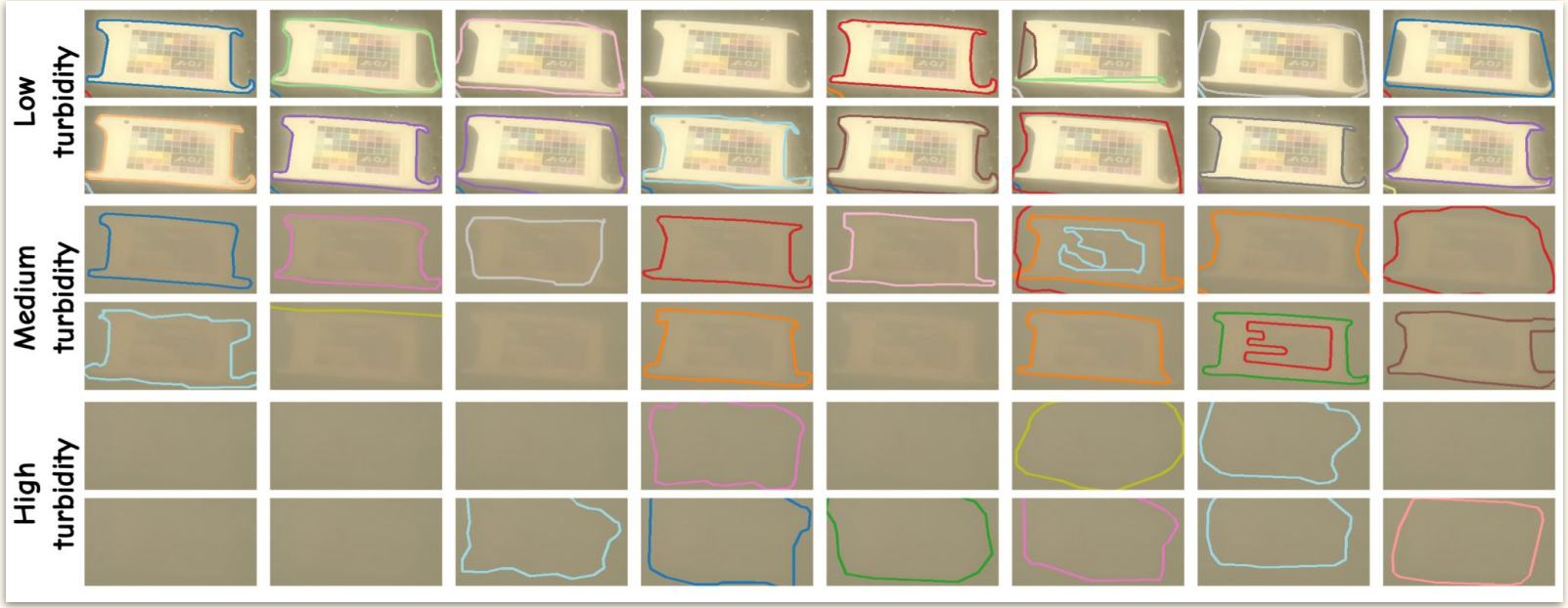
A concrete example: TUBcertainty

What happens when you ask 100+ people to annotate objects in underwater images?



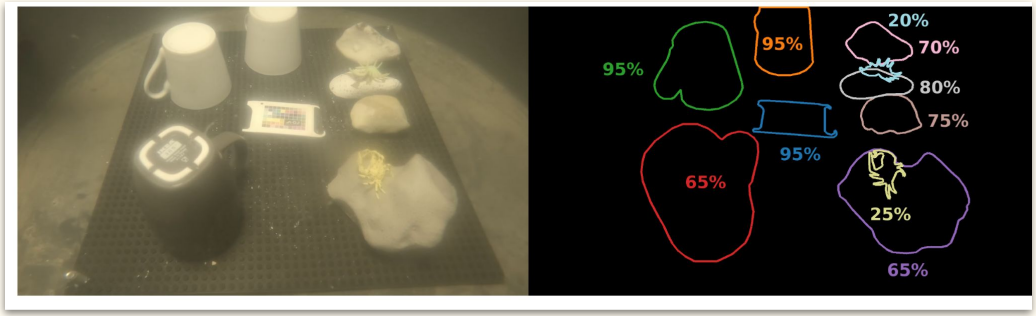
A Multi-Annotator Study of Segmentation Noise and Uncertainty in Turbid Underwater Images ECCVW 2026

A concrete example: TUBcertainty

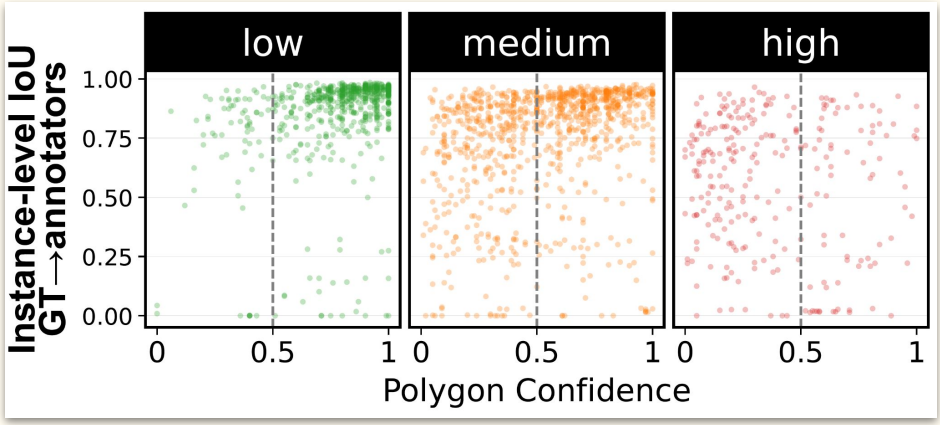


A Multi-Annotator Study of Segmentation Noise and Uncertainty in Turbid Underwater Images ECCVW 2026

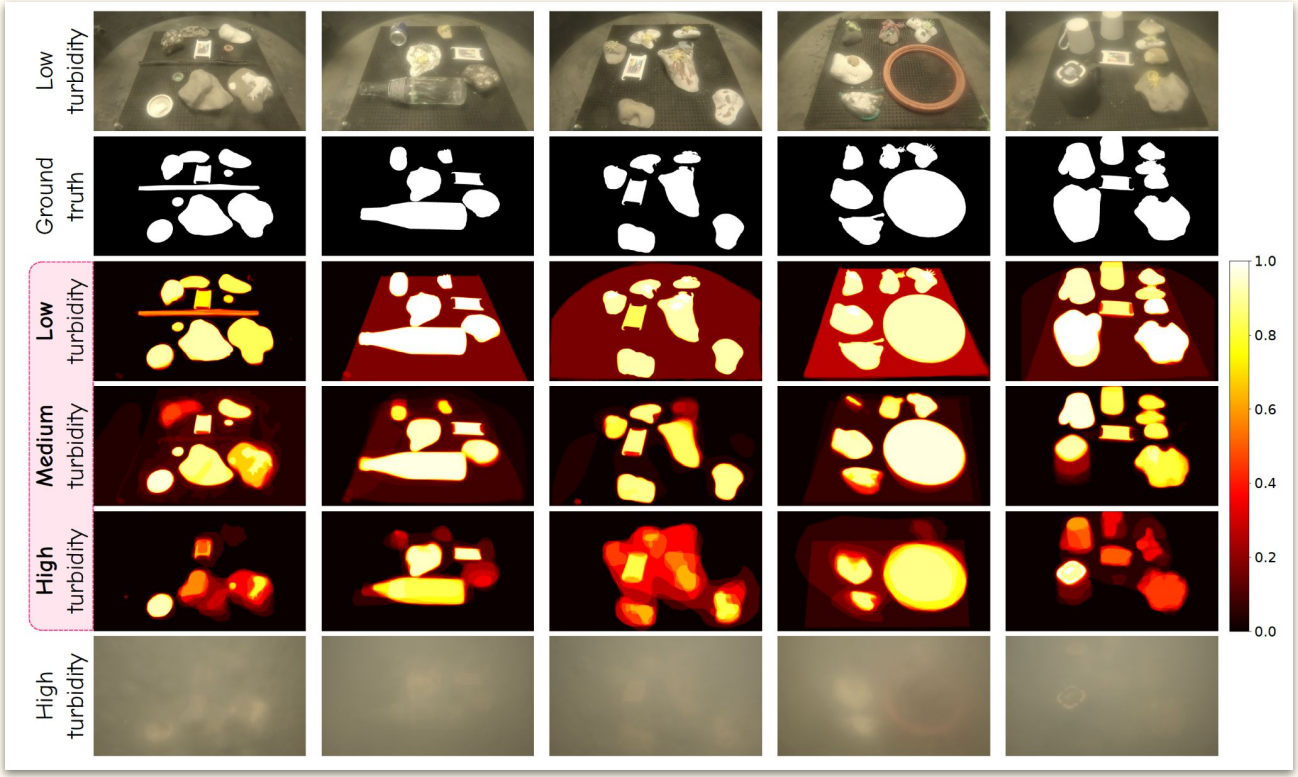
A concrete example: TUBcertainty



Human annotators are not necessarily well-calibrated, and confidence can also be a noisy signal!
Prior knowledge also heavily influences self-reported confidence



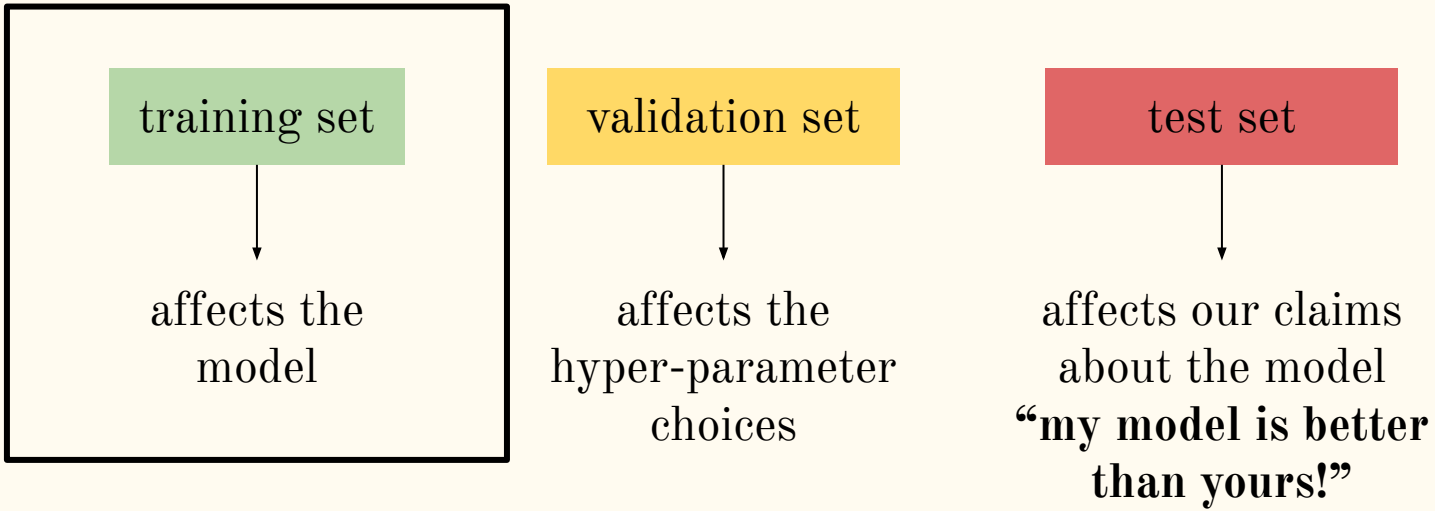
A concrete example: TUBcertainty



A Multi-Annotator Study of Segmentation Noise and Uncertainty in Turbid Underwater Images ECCVW 2026

How do imperfect labels affect model performance and post-hoc tasks

Implications of labeling errors



Are classifiers robust to label noise?

TL;DR **it depends**, and you will find seemingly contradictory claims in the literature.

- it depends not only on the noise rate but also on the **number of clean samples** (high robustness can be achieved if the number of clean labels is high enough, and noisy labels are added instead of **replacing** clean labels with noisy ones). **The critical amount of clean data increases as the noise level rises.**
- it depends on the **noise model**. Uniform label noise is generally less detrimental but also the least realistic.
- it depends **when training stops**: there seems to be an "early learning phase" during which the network correctly learns clean label patterns, before overfitting to mislabeled examples as training progresses

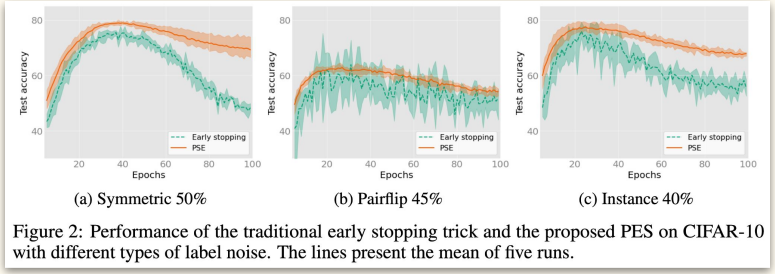
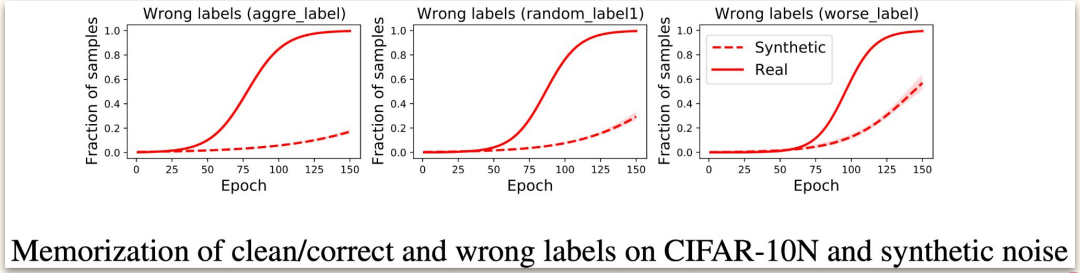


Figure 2: Performance of the traditional early stopping trick and the proposed PES on CIFAR-10 with different types of label noise. The lines present the mean of five runs.



Memorization of clean/correct and wrong labels on CIFAR-10N and synthetic noise

Are classifiers robust to label noise?

TL;DR **it depends**, and you will find seemingly contradictory claims in the literature.

- starting from a pre-trained model and fine-tuning (rather than training from scratch) helps a lot
- segmentation: “each architecture was very robust to non-boundary-localized errors, suggesting that boundary-localized errors are fundamentally different and more challenging problem than random label errors”

Problem: these studies require a “clean” version of the training set for comparison. Thus many rely on re-annotated toy datasets, or on synthetically injected label noise.

[1705.10694] Deep Learning is Robust to Massive Label Noise 2018
Beyond Synthetic Noise: Deep Learning on Controlled Noisy Labels ICML 2020
[1806.04618] Imperfect Segmentation Labels: How Much Do They Matter? LABELS, 2018
[2106.15853] Understanding and Improving Early Stopping for Learning with Noisy Labels NeurIPS 2021
Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations ICLR 2022

What about downstream tasks?

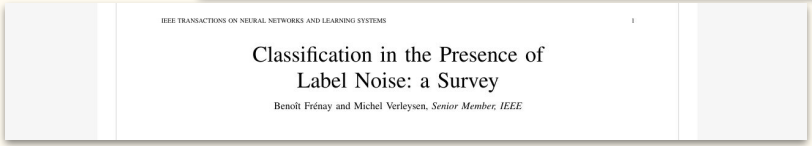
- investigating and improving robustness to label noise is an active field of research


A Survey on Classifying Big Data with Label Noise

JUSTIN M. JOHNSON and TAGHI M. KHOSHGOFTAAR, Florida Atlantic University, USA


Noisy Label Processing for Classification: A Survey

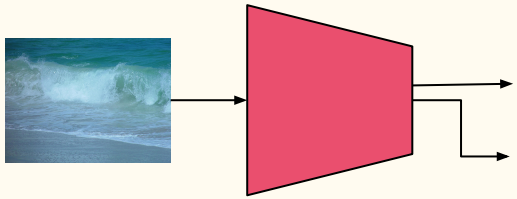
Mengting Li[a], Chuang Zhu[b,**]


IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS 1

Classification in the Presence of Label Noise: a Survey

Benoît Frénay and Michel Verleysen, *Senior Member, IEEE*

- however, the effects of training-time label noise on downstream post-hoc tasks is not well-explored (e.g. OOD detection, XAI)

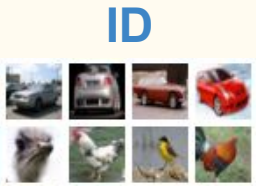


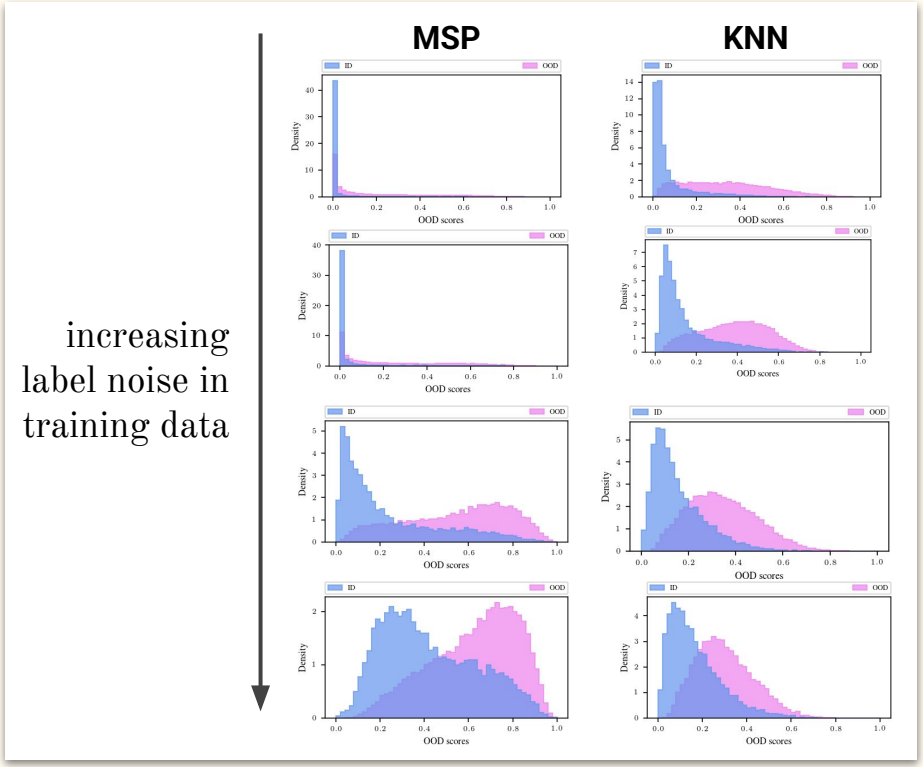
class prediction

is this a known or unknown class?

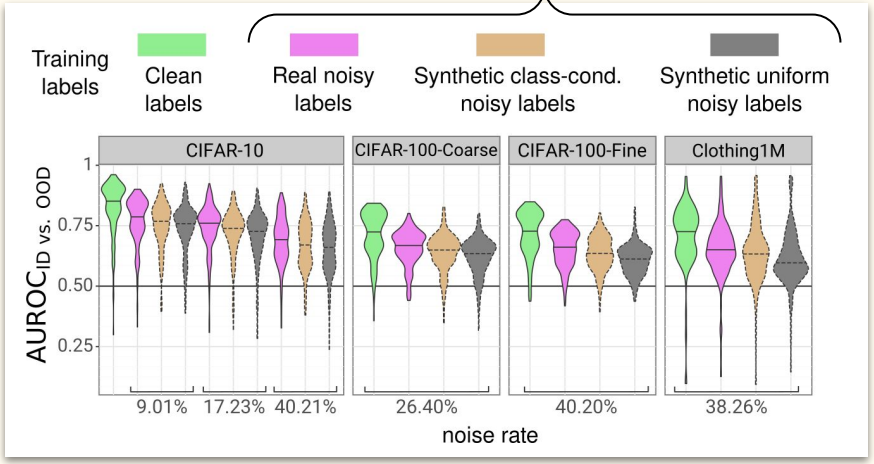
which parts of the input are decisive for the prediction?

Impact of label noise on OOD detection





different noise models



[2404.01775] A noisy elephant in the room: Is your out-of-distribution detector robust to label noise?
CVPR, 2024.

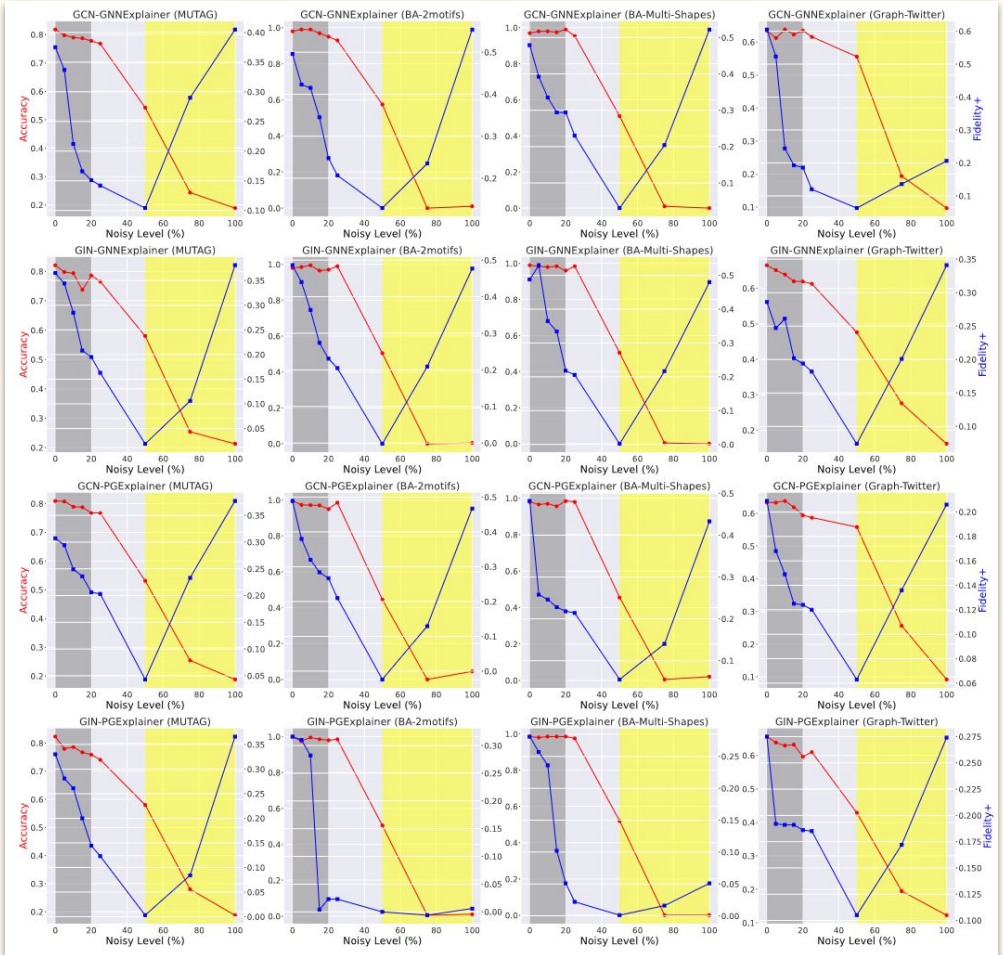


Figure 1: The performance of GNN models and explainers of different label noisy levels.

Impact of label noise on explanations

Very little work on this...

The little work on this suggests that there might be non-linear & counter-intuitive effects

[2309.01706] On the Robustness of Post-hoc GNN Explainers to Label Noise
Learning on Graphs, 2023.

Modeling/learning with data uncertainty

How to deal with imperfect/uncertain labels

Common things to do:

- disregard low-quality/low-confidence samples
- take the majority vote across multiple annotators
- ignore the problem and hope for the best

This disregards valuable information and/or leads to weaker models!
It's better to embrace imperfection

- heteroscedastic aleatoric uncertainty estimation
- probabilistic embeddings
- noisy label learning
- multi-annotator label learning / learning from disagreement

Heteroscedastic classifiers

Overall idea: to model input-dependent aleatoric uncertainty, place a Gaussian distribution over the logits, predict the (co-)variance alongside the mean, then pass samples through softmax.

Early works assume a **diagonal covariance** (A,B), while more recent works (C,D) **approximate the full covariance matrix** to model noise/overlap between specific class pairs.

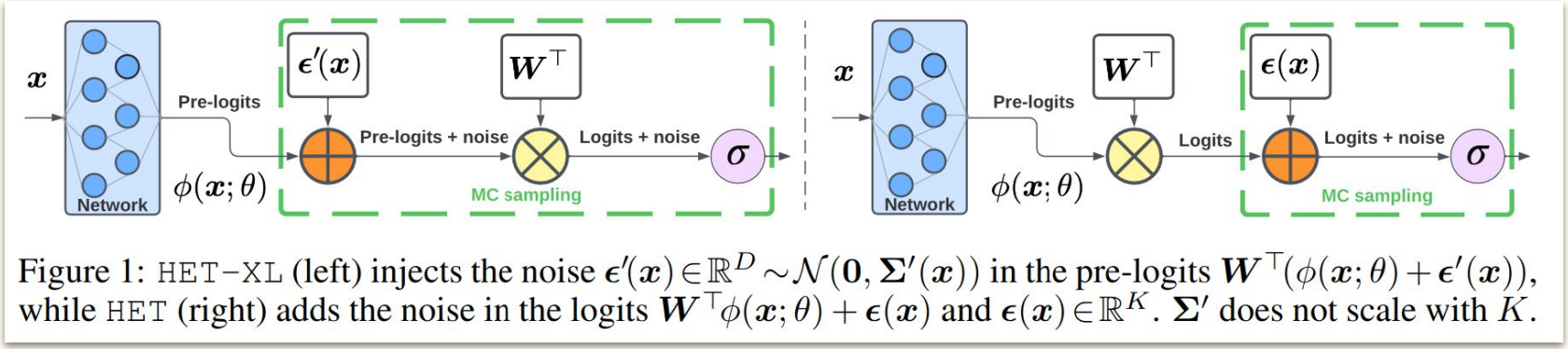


Figure 1: HET-XL (left) injects the noise $\boldsymbol{\epsilon}'(\boldsymbol{x}) \in \mathbb{R}^D \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}'(\boldsymbol{x}))$ in the pre-logits $\boldsymbol{W}^\top(\phi(\boldsymbol{x};\theta) + \boldsymbol{\epsilon}'(\boldsymbol{x}))$, while HET (right) adds the noise in the logits $\boldsymbol{W}^\top\phi(\boldsymbol{x};\theta) + \boldsymbol{\epsilon}(\boldsymbol{x})$ and $\boldsymbol{\epsilon}(\boldsymbol{x}) \in \mathbb{R}^K$. $\boldsymbol{\Sigma}'$ does not scale with K.

A [1703.04977] What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NeurIPS, 2017.
B [2003.06778] A Simple Probabilistic Method for Deep Classification under Input-Dependent Label Noise 2020
C [2105.10305] Correlated Input-Dependent Label Noise in Large-Scale Image Classification CVPR 2021
D [2301.12860] Massively Scaling Heteroscedastic Classifiers ICLR, 2023.

Heteroscedastic classifiers

standard softmax

Method	Webvision			ILSVRC12		
	Top-1 Acc	Top-5 Acc	NLL	Top-1 Acc	Top-5 Acc	NLL
Homoscedastic	76.1† (±0.07)	91.2† (±0.07)	1.03† (±0.006)	67.0† (±0.08)	86.0† (±0.11)	1.49† (±0.008)
Het. Diag [9]	76.2† (±0.15)	91.4† (±0.08)	1.01† (±0.002)	67.3† (±0.10)	86.1† (±0.07)	1.47† (±0.004)
Het. Full (ours)	**76.6** (±0.13)	**92.1** (±0.09)	**0.98** (±0.004)	**68.6** (±0.17)	**87.1** (±0.13)	**1.41** (±0.010)

looking at the predicted covariance across different samples can give insight into the **most easily confused classes** in a dataset

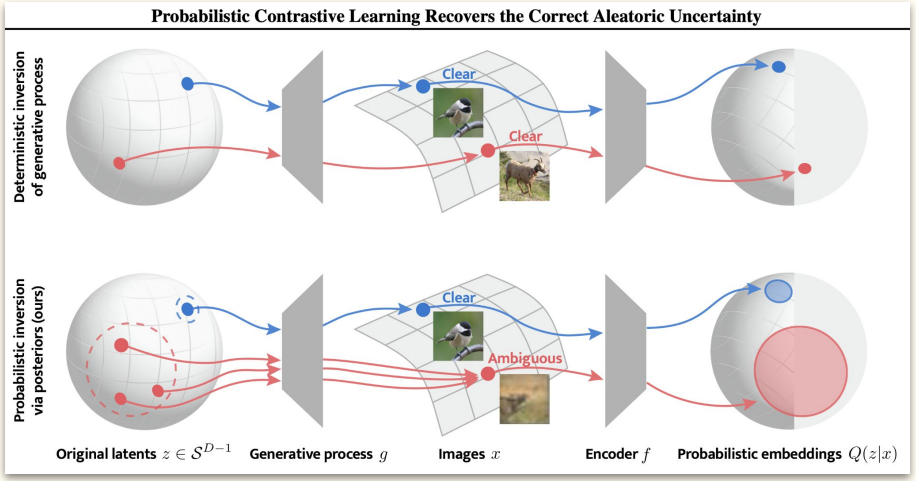
Class A	Class B	Avg. Cov.
partridge	ruffed grouse, partridge, Bonasa umbellus	-0.46
projectile, missile	missile	-0.44
maillot, tank suit	maillot	-0.42
screen, CRT screen	monitor	-0.37
tape player	cassette player	-0.34
Welsh springer spaniel	Blenheim spaniel	0.28
standard schnauzer	miniature schnauzer	0.27
French bulldog	Boston bull, Boston terrier	0.27
EntleBucher	standard schnauzer	0.26
tennis ball	racket, racquet	-0.26

Table 4: Class pairs with the top-10 absolute covariance, averaged over the ILSVRC12 validation set.

[2105.10305] Correlated Input-Dependent Label Noise in Large-Scale Image Classification CVPR 2021

Probabilistic embeddings

By default, metric learning methods represent the input **as a single point in the embedding space**. This disregards the fact that the similarity of two samples may be inherently uncertain.



[1810.00319] Modeling Uncertainty with Hedged Instance Embedding ICLR 2019
[1904.09658] Probabilistic Face Embeddings ICCV 2019
[2011.12663] Bayesian Triplet Loss: Uncertainty Quantification in Image Retrieval ICCV 2021
[2101.05068] Probabilistic Embeddings for Cross-Modal Retrieval CVPR 2021
[2204.03946] Probabilistic Representations for Video Contrastive Learning CVPR 2022
Maximum Entropy Information Bottleneck for Uncertainty-Aware Stochastic Embedding CVPRW 2023
[2302.02865] Probabilistic Contrastive Learning Recovers the Correct Aleatoric Uncertainty of Ambiguous Inputs ICML 2023
URL: A Representation Learning Benchmark for Transferable Uncertainty Estimates | OpenReview NeurIPS 2023
Probabilistic Embeddings for Frozen Vision-Language Models UAI 2025

Probabilistic embeddings

Overall idea:

- the encoder outputs the parameters of a distribution (e.g. Gaussian, von Mises-Fisher, Mixture-of-Gaussians)
- adapt the contrastive loss, either via sampling or probabilistic distance, with a regularizer to prevent collapse
- at test-time, the spread can be used as a measure of aleatoric uncertainty for downstream tasks (e.g. retrieval)
- recent work has extended this to frozen embeddings & VLMs

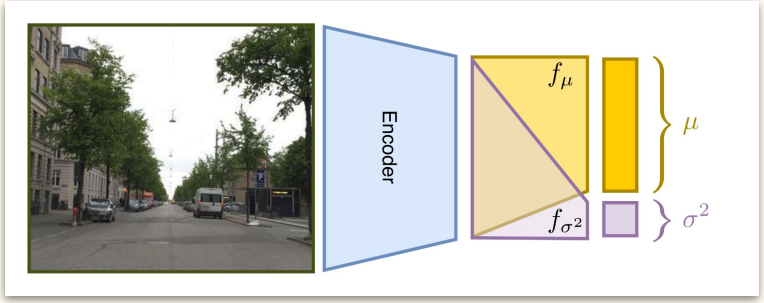
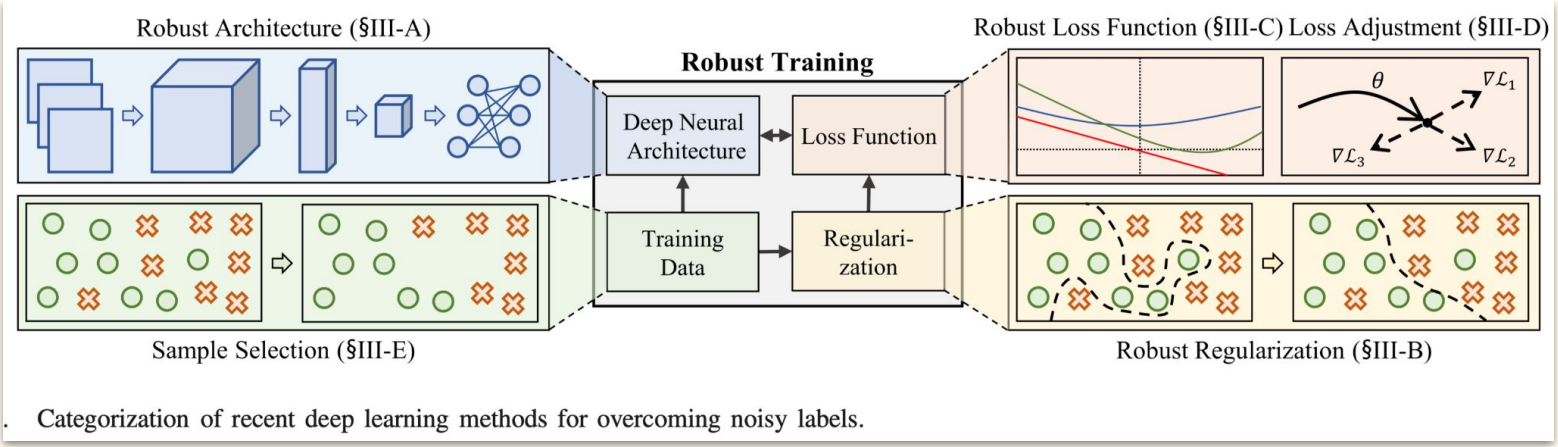


Figure 6: Six query images (top) and their NN (bottom), all true positives. Our Bayesian model assigns low confidence to images with multiple birds or where birds blend in with their surroundings. Birds with distinguishable patterns or colors have high confidence.

Noisy label learning

Many different directions for dealing with labeling errors in the training set. Most methods rely on some underlying assumptions about the noise model (typically instance-independence) in order to e.g. estimate the noise transition matrix.



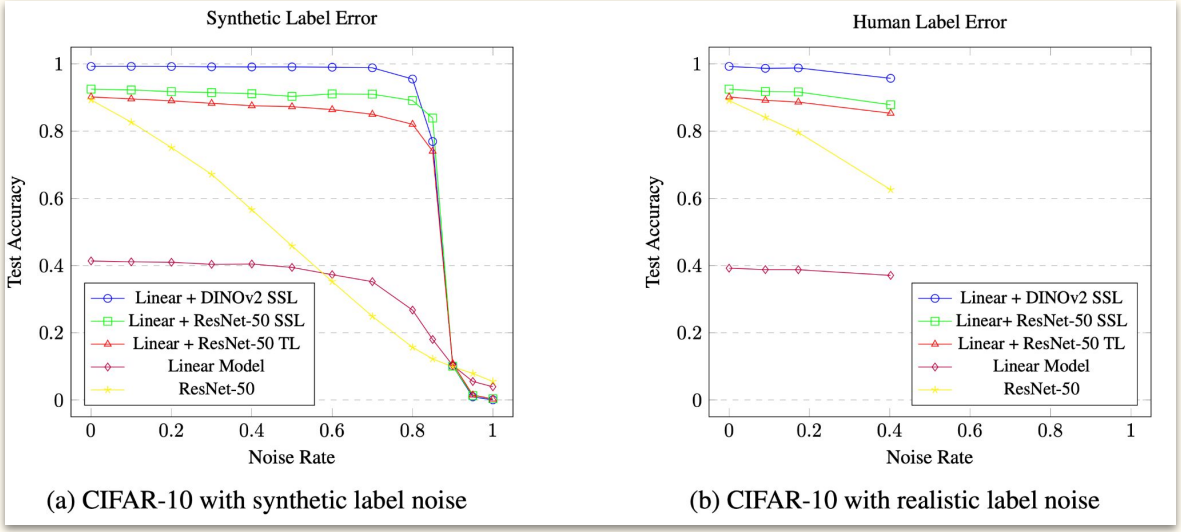
. Categorization of recent deep learning methods for overcoming noisy labels.

[2007.08199] Learning from Noisy Labels with Deep Neural Networks: A Survey TNNL 2022

Noisy label learning

1. Perform feature extraction using any method (e.g., transfer learning or self-supervised learning) that does not require labels.
2. Learn a simple classifier (e.g., a linear classifier) on top of these extracted features, using the noisily labelled data, in a noise-ignorant way.

Forget about fine-tuning or training from scratch, **just use DINO + linear probe?**

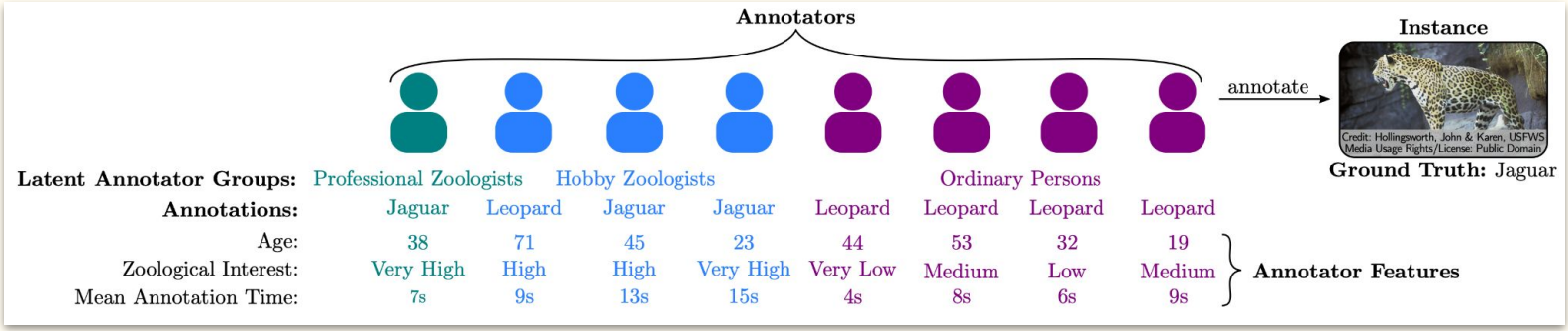


(a) CIFAR-10 with synthetic label noise

(b) CIFAR-10 with realistic label noise

[2411.00079] Label Noise: Ignorance Is Bliss NeurIPS 2024

Multi-annotator label learning



How can we leverage the wisdom of the crowd during training?

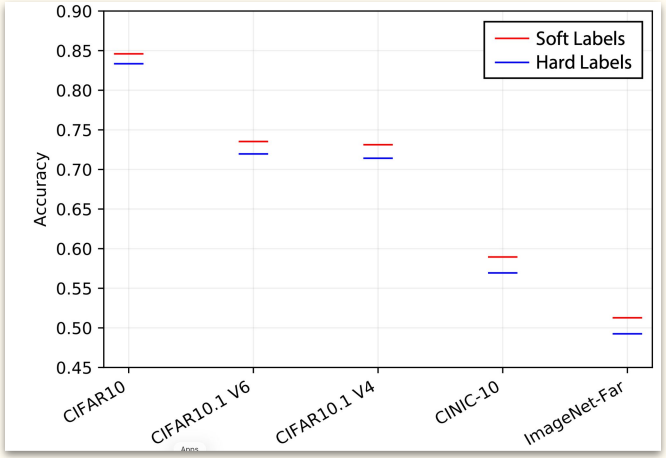
- **Two-stage approaches**: First **estimate a single label** from multiple annotator inputs (e.g. smart aggregation, soft labeling) and then train with standard MLE
- **One-stage approaches**: Train on the **raw noisy labels directly** and explicitly model the annotator disagreement during training.

Learning from Disagreement: : A Survey: Journal of Artificial Intelligence Research Journal of Artificial Intelligence Research, 2022
[2304.02539] Multi-annotator Deep Learning: A Probabilistic Framework for Classification TMLR, 2023
[2407.20950] dopanim: A Dataset of Doppelganger Animals with Noisy Annotations from Multiple Humans NeurIPS 2024

Multi-annotator label learning

How can we leverage the wisdom of the crowd during training?

- **Two-stage approaches:** Soft labeling is an easy and strong baseline!



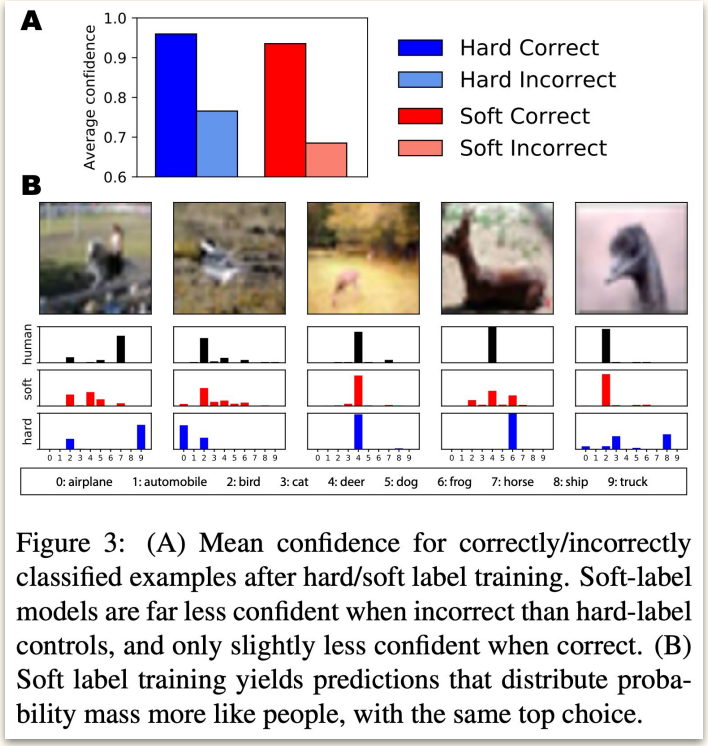


Figure 3: (A) Mean confidence for correctly/incorrectly classified examples after hard/soft label training. Soft-label models are far less confident when incorrect than hard-label controls, and only slightly less confident when correct. (B) Soft label training yields predictions that distribute probability mass more like people, with the same top choice.

[1908.07086] Human uncertainty makes classification more robust ICCV 2019

Multi-annotator label learning

How can we leverage the wisdom of the crowd during training?

- **One-stage approaches:** These usually tackle the problem by modeling class-dependent or instance-dependent annotator performances

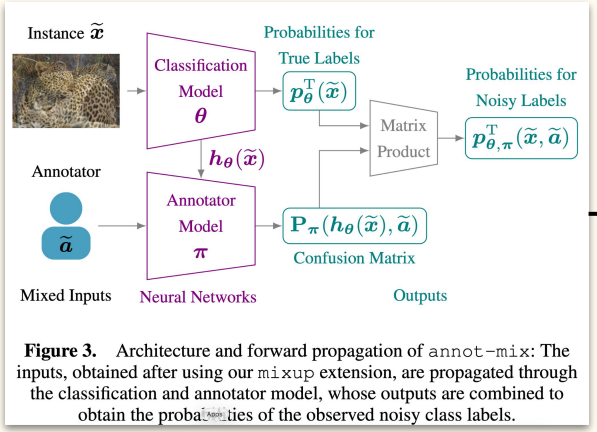


Figure 3. Architecture and forward propagation of `annot-mix`: The inputs, obtained after using our `mixup` extension, are propagated through the classification and annotator model, whose outputs are combined to obtain the probabilities of the observed noisy class labels.

majority vote

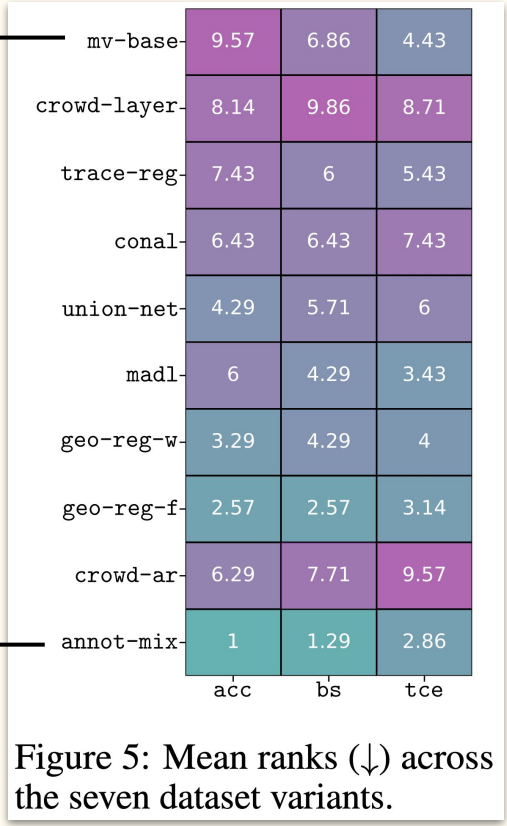


Figure 5: Mean ranks (↓) across the seven dataset variants.

[2407.20950] dopanim: A Dataset of Doppelganger Animals with Noisy Annotations from Multiple Humans NeurIPS 202
[2405.03386] Annot-Mix: Learning with Noisy Class Labels from Multiple Annotators via a Mixup Extension ECAI 2024

Wisdom from NLP: dataset cartography

We define **confidence** as the mean model probability of the true label (y_i^*) across epochs:

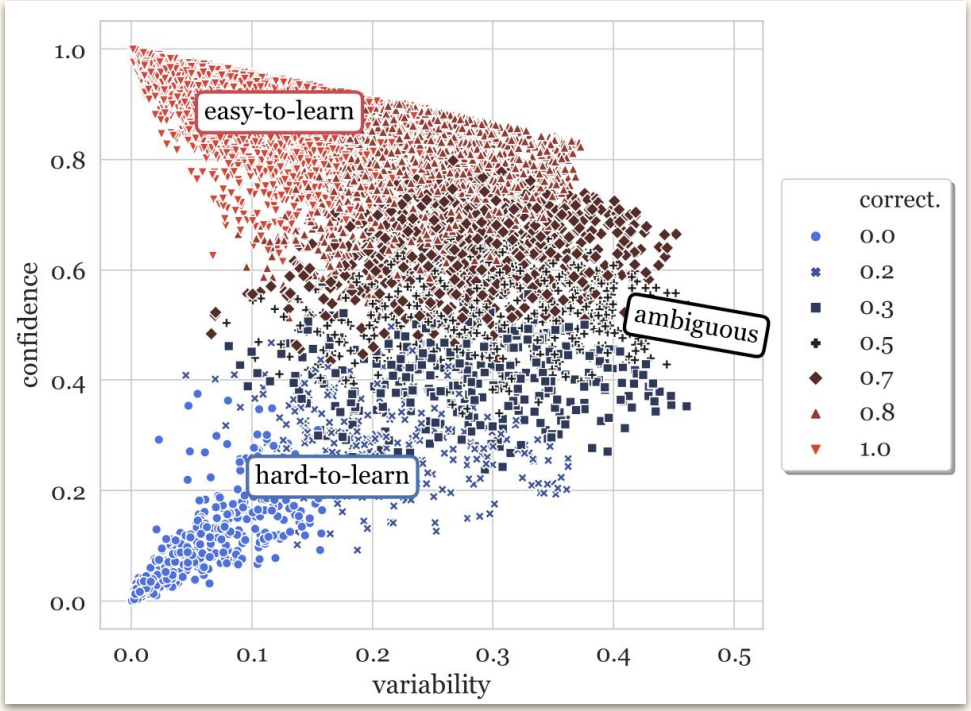
$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\boldsymbol{\theta}^{(e)}}(y_i^* \mid \boldsymbol{x}_i)$$

where $p_{\boldsymbol{\theta}^{(e)}}$ denotes the model's probability with parameters $\boldsymbol{\theta}^{(e)}$ at the end of the e^{th} epoch.[5] In

Lastly, we also consider **variability**, which measures the spread of $p_{\boldsymbol{\theta}^{(e)}}(y_i^* \mid \boldsymbol{x}_i)$ across epochs, using the standard deviation:

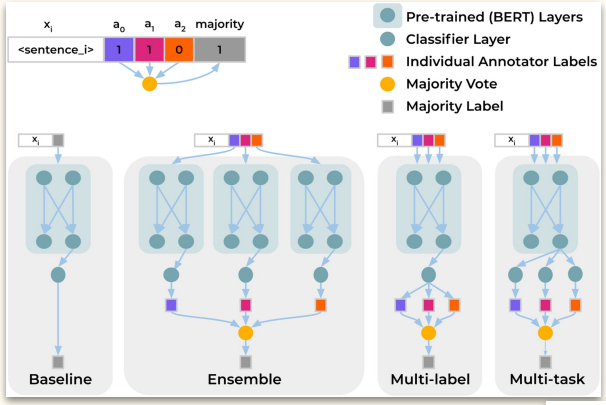
$$\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}\left(p_{\boldsymbol{\theta}^{(e)}}(y_i^* \mid \boldsymbol{x}_i) - \hat{\mu}_i\right)^2}{E}}$$

Note that **variability** also depends on the gold label, y_i^*. A given instance to which the model assigns the same label consistently (whether accurately or not) will have low **variability**; one which the model is indecisive about across training, will have high **variability**.



[2009.10795] Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics EMNLP, 2020.

Wisdom from NLP: modeling individual annotators



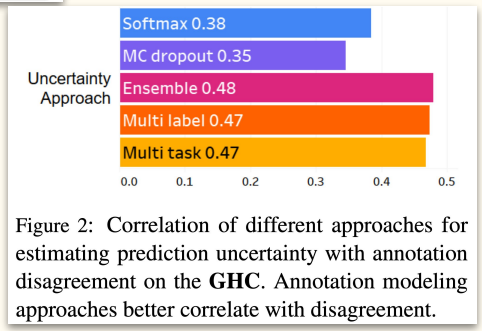


Figure 2: Correlation of different approaches for estimating prediction uncertainty with annotation disagreement on the **GHC**. Annotation modeling approaches better correlate with disagreement.

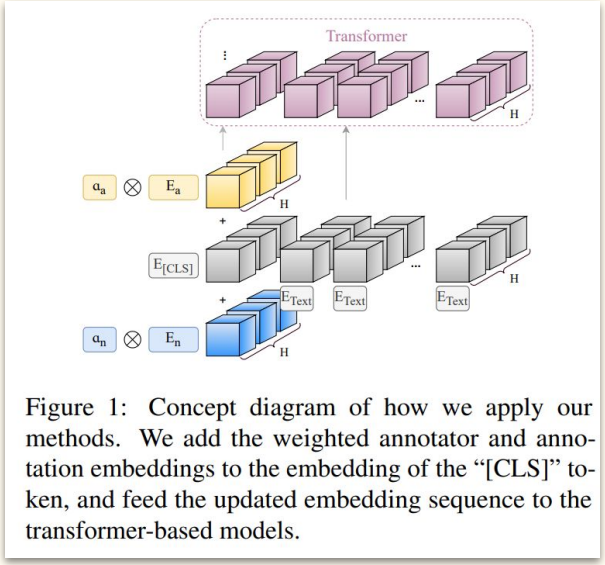


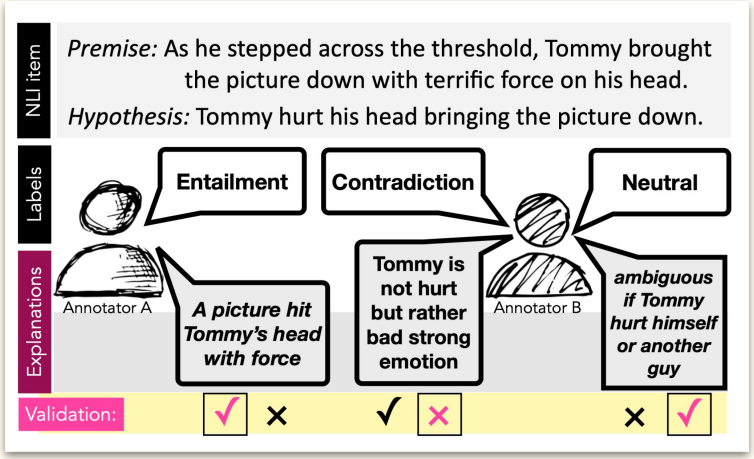
Figure 1: Concept diagram of how we apply our methods. We add the weighted annotator and annotation embeddings to the embedding of the "[CLS]" token, and feed the updated embedding sequence to the transformer-based models.

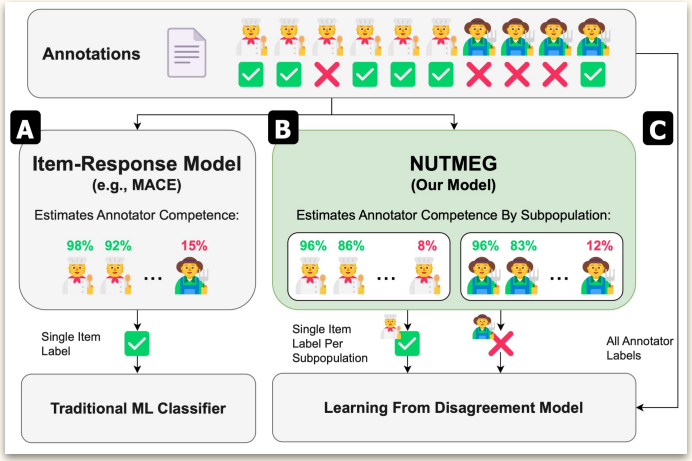
[2110.05719] Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations Transactions of the ACL, 2022

You Are What You Annotate: Towards Better Models through Annotator Representations - ACL Anthology EMNLP, 2023

Wisdom from NLP: separating error from variation

Inter-annotator disagreement can arise due to many different reasons e.g. different expertise, backgrounds, motivations, unintentional mistakes and adversarial behaviour. **Some disagreements are valid variation while some are undesirable errors.** These are usually studied in isolation, but it's important to be able to disentangle them in practice!





VariErr NLI: Separating Annotation Error from Human Label Variation - ACL Anthology ACL 2024
NUTMEG: Separating Signal From Noise in Annotator Disagreement, EMNLP 2025

How do we evaluate aleatoric uncertainty estimates?

Evaluating epistemic vs. aleatoric uncertainty

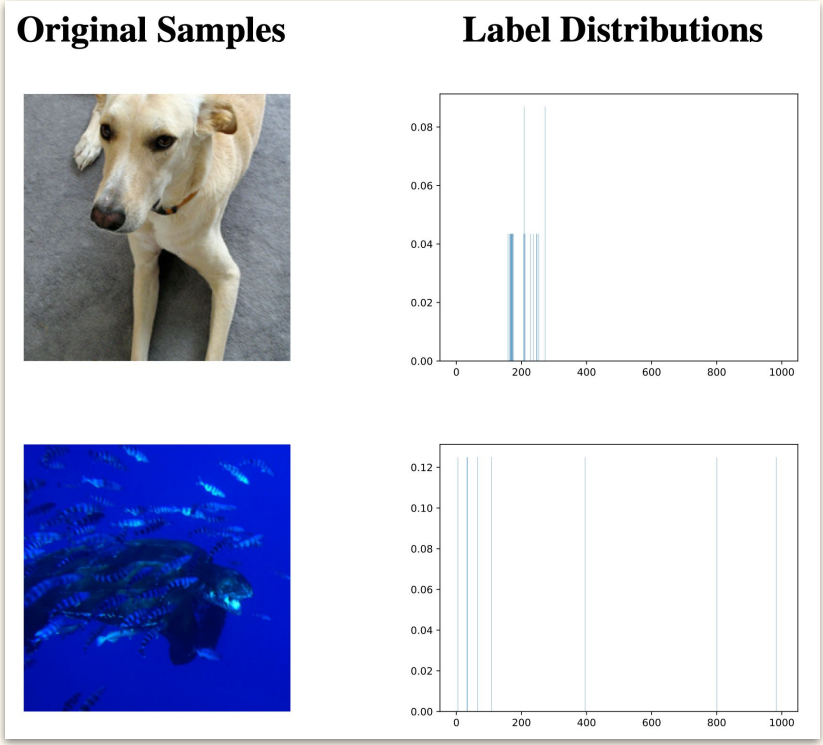
When evaluating **epistemic or predictive uncertainty**, we use the model correctness as GT: if the model is **incorrect**, uncertainty should be high and vice-versa.

└ compared to a **single GT label**

However, the **GT aleatoric uncertainty is independent of the model prediction**: even if the model is correct, uncertainty should be high if the input is ambiguous.

- ➔ Problem: we usually don't have access to the true $\mathbf{P(Y \mid X = x)}$. We usually only have a single label per image with no GT measure of label uncertainty.
- ➔ Best-case scenario: we have **samples from $\mathbf{P(Y \mid X = x)}$** (e.g. multi-annotator labels in the test set, the more the better), which we can use as a GT proxy.
- ➔ Otherwise, many works simply evaluate classification accuracy or NLL, to show that modeling aleatoric uncertainty is beneficial.

Directly evaluating aleatoric uncertainty

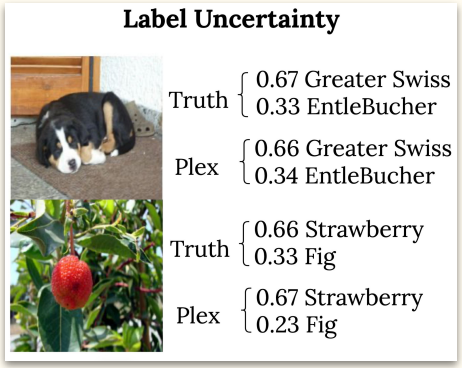


(A) [2006.07159] Are we done with ImageNet? 2020

Examples:

A. use the entropy of the soft-label distributions per image as GT aleatoric uncertainties. Then calculate the rank **correlation between the methods' uncertainty estimates and the GT label entropies** across all images.

B. or directly measure the KL-divergence between the GT label distribution and estimated one



(B) [2207.07411] Plex: Towards Reliability using Pretrained Large Model Extensions 2022

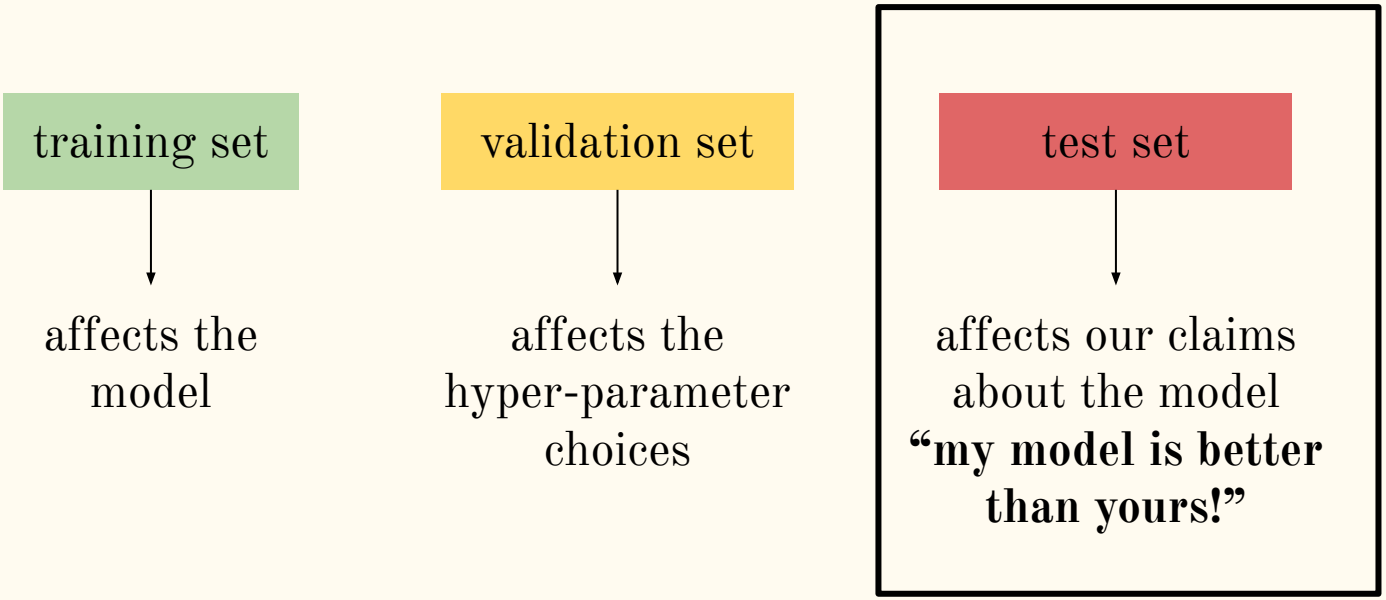
Directly evaluating aleatoric uncertainty

[2402.19460] Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks
NeurIPS, 2024

- "While epistemic uncertainty is widely evaluated on the OOD detection proxy task, **aleatoric uncertainty still lacks a standardized testing protocol**. The current approaches seem to converge to **soft labels**, but nuances in how they are collected still need discussion (compare, for example, CIFAR-10H to CIFAR-10S and CIFAR-10N).
- An increasing number of uncertainty quantification approaches compare to such **human GT notions of aleatoric uncertainty**, indicating the interest in the field.
- Our benchmark shows that no method can give highly accurate aleatoric uncertainty estimates yet, stressing the **need for benchmarks, methods, and training resources to develop along.**"

Labels used for evaluation

Implications of labeling errors



Labeling errors in benchmarks

Undocumented errors have been found in many widely used ML benchmarks, even “easy”/toy datasets - this includes the test sets used for evaluation!

Object365

Open Images

MNIST

Pascal VOC

ADE20K

KITTI

ImageNet

COCO

Caltech-256

and many more!

CIFAR

Cityscapes

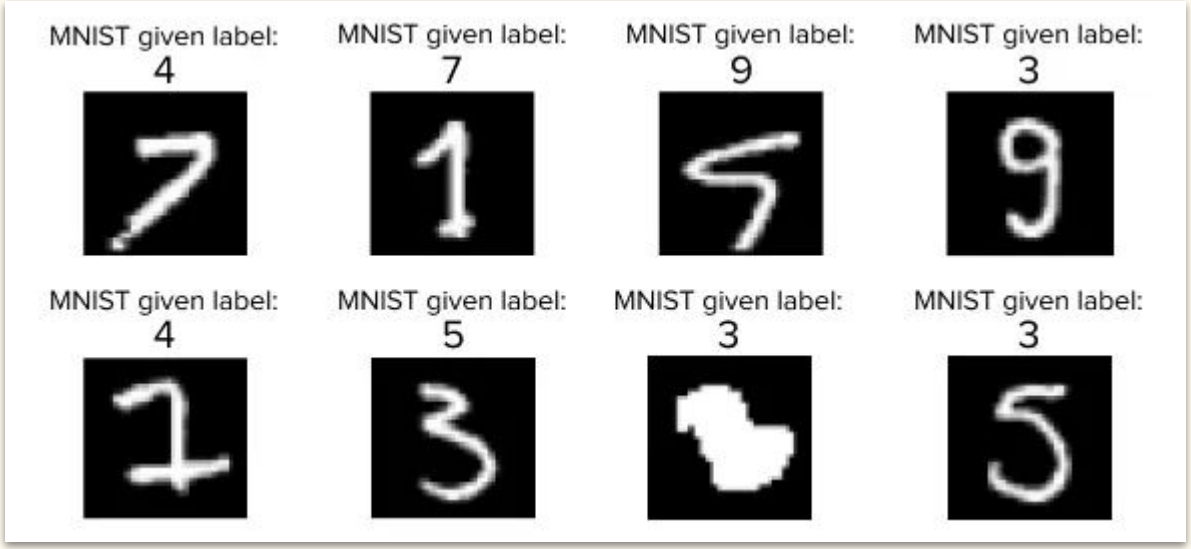
LiveCell

QuickDraw

[1911.00068] Confident Learning: Estimating Uncertainty in Dataset Labels Journal of Artificial Intelligence Research, 2021
[2103.14749] Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks NeurIPS, 2021.
[2508.17930] Learning to Detect Label Errors by Making Them: A Method for Segmentation and Object Detection Datasets
Quality over quantity: a data-centric survey of annotation errors in object detection datasets Artificial Intelligence Review, 2026

Labeling errors in benchmarks

Undocumented errors have been found in many widely used ML benchmarks, even “easy”/toy datasets - this includes the test sets used for evaluation!



“To our surprise, the original, unperturbed MNIST dataset, which is predominately assumed error-free, contains blatant label errors”

[1911.00068] Confident Learning: Estimating Uncertainty in Dataset Labels
Journal of Artificial Intelligence Research, 2021

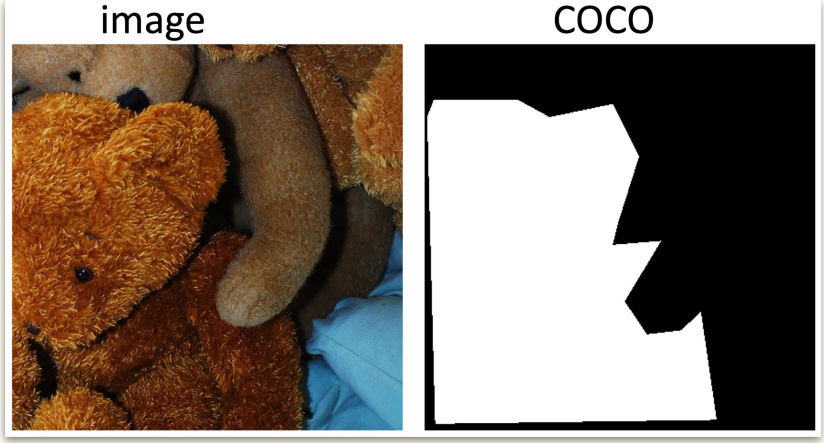
Labeling errors in benchmarks

Undocumented errors have been found in many widely used ML benchmarks, even "easy"/toy datasets - this includes the test sets used for evaluation!

"COCO's early design inevitably encompassed certain annotation biases, including issues like **imprecise object boundaries and incorrect class labels.**

- COCO's annotation is neither exhaustive nor consistent
- instances of incomplete labeling are commonplace
- discrepancies between semantic, instance, and panoptic annotations within the dataset present challenges in developing a comprehensive segmentation model.

The rapid evolution of model architectures has led to a **performance plateau on the COCO benchmark**. This stagnation suggests a potential **overfitting to the dataset's specific characteristics**, raising concerns about the models' applicability to real-world data."



COCONut: Modernizing COCO Segmentation CVPR, 2024

Labeling errors in benchmarks

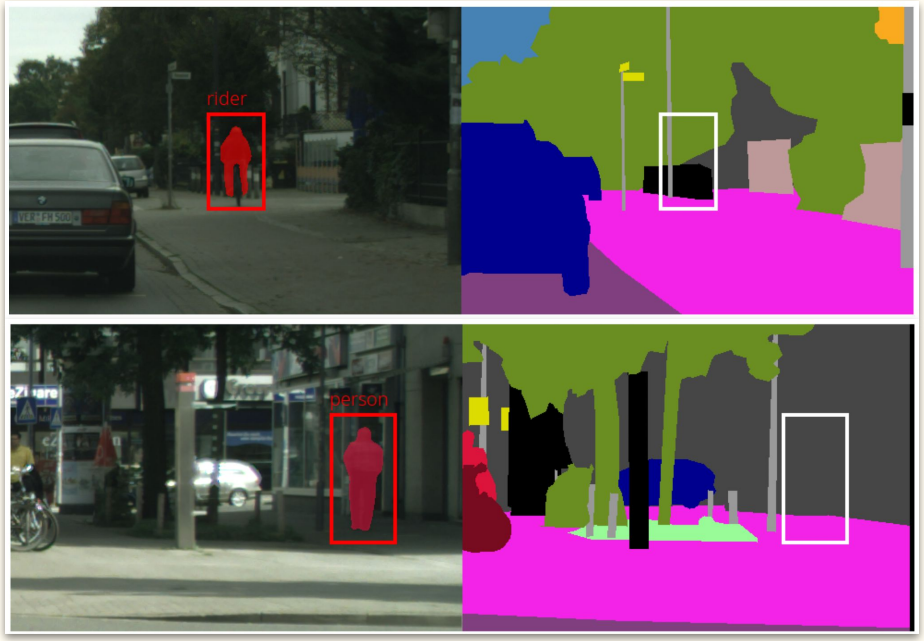
Undocumented errors have been found in many widely used ML benchmarks, even “easy”/toy datasets - this includes the test sets used for evaluation!

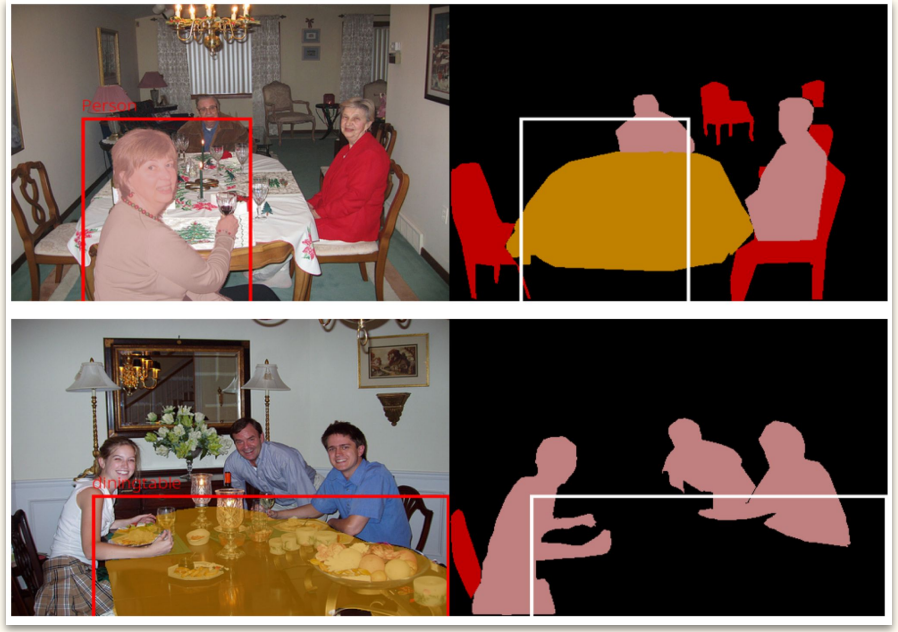
“we acquired high-quality annotations based on which we identified **at least 24.6% missing and inaccurate pedestrian annotations** in the original KITTI dataset.” [2508.06556] From Label Error Detection to Correction: A Modular Framework and Benchmark for Object Detection Datasets 2026

Labeling errors in benchmarks

Cityscapes

PascalVOC

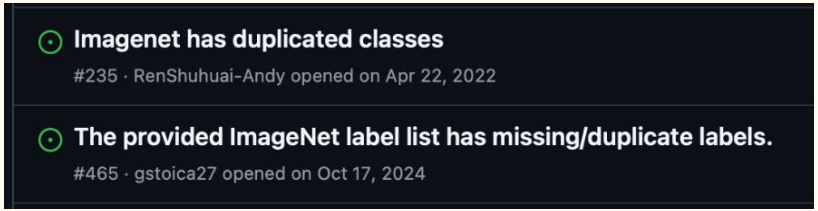




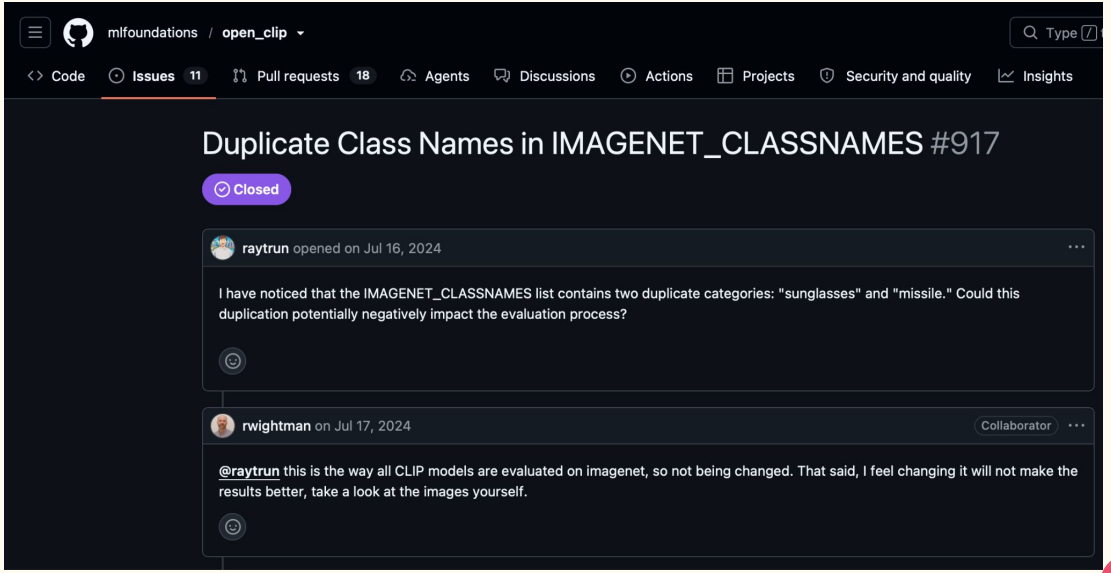
[2207.06104] Automated Detection of Label Errors in Semantic Segmentation Datasets via Deep Learning and Uncertainty Quantification WACV, 2023

Another slightly frustrating example

github.com/openai/clip

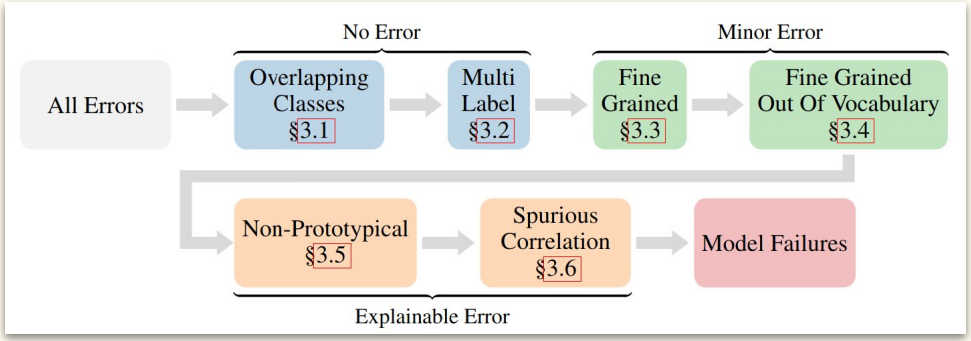


github.com/mlfoundations/open_clip/



Labeling errors in benchmarks

Do small performance increases indicate true progress or are we over-fitting to the idiosyncrasies of the dataset? Model prediction may sometimes end up "more correct" than the GT label...



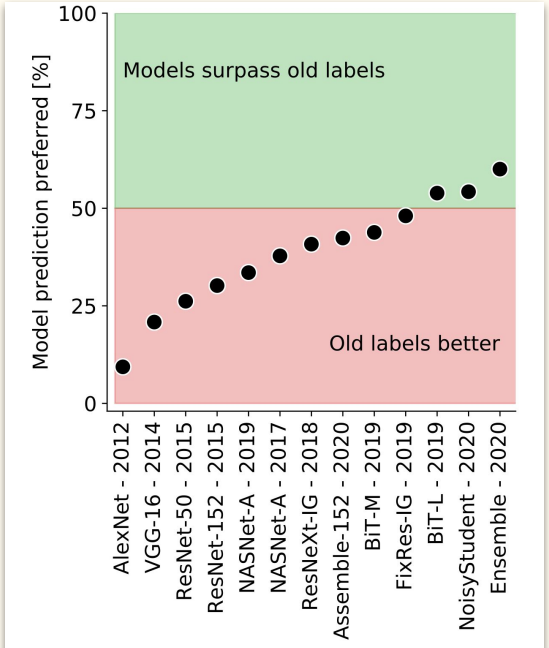


Figure 1: When presented with a model's prediction and the original ImageNet label, human annotators now prefer model predictions on average (Section 4). Nevertheless, there remains considerable progress to be made before fully capturing human preferences.

[2401.02430] Automated Classification of Model Errors on ImageNet NeurIPS, 2023.
[2006.07159] Are we done with ImageNet? 2020

Labeling errors in benchmarks

Do small performance increases indicate true progress or are we over-fitting to the idiosyncrasies of the dataset?

- note that this also applies to other parallel tasks that also rely on GT labels for evaluation!

  e.g. **if you have labeling errors in the test set, this will also unfairly penalize uncertainty estimation metrics** (ECE, Brier, AUROC, etc)

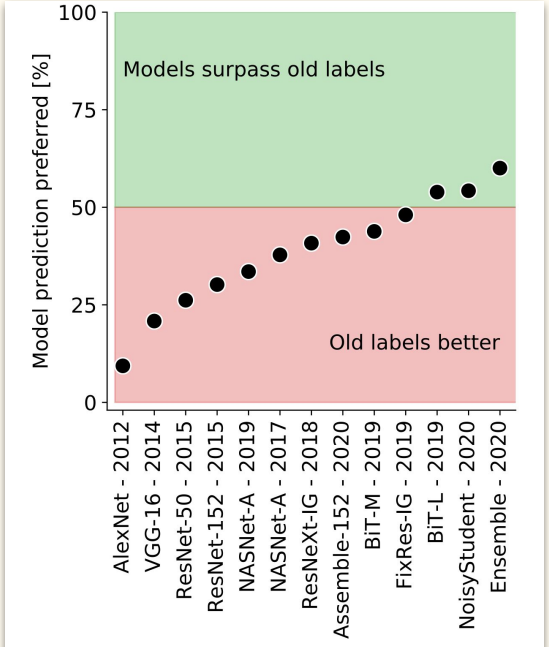
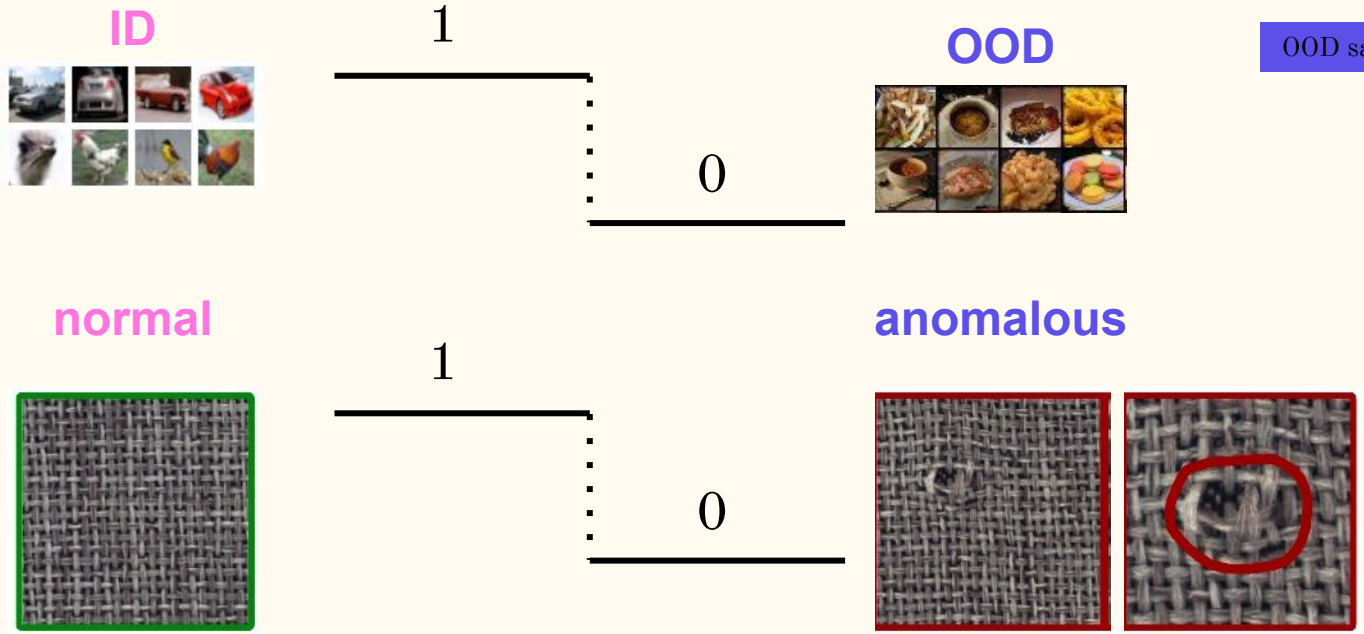
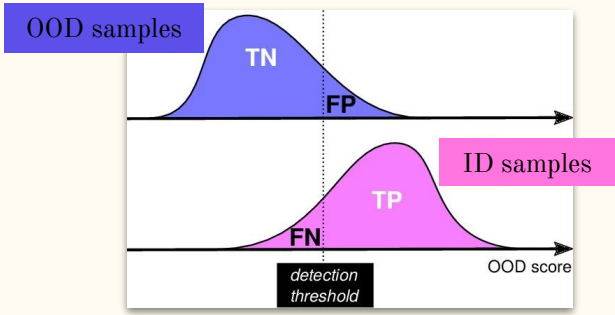


Figure 1: When presented with a model's prediction and the original ImageNet label, human annotators now prefer model predictions on average (Section 4). Nevertheless, there remains considerable progress to be made before fully capturing human preferences.

Labels for OOD detection and anomaly detection

In OOD & anomaly detection we adopt the following binary evaluation setting:



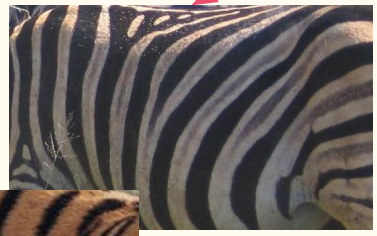


Labels in OOD detection

includes **ID classes:**
zebra, tiger, computer mouse, beer bottle

Leaderboard: ImageNet-1K

- **Near(Hard)-OOD:** SSB-hard, NINCO
- **Far(Easy)-OOD:** iNaturalist, Textures, OpenImage-O

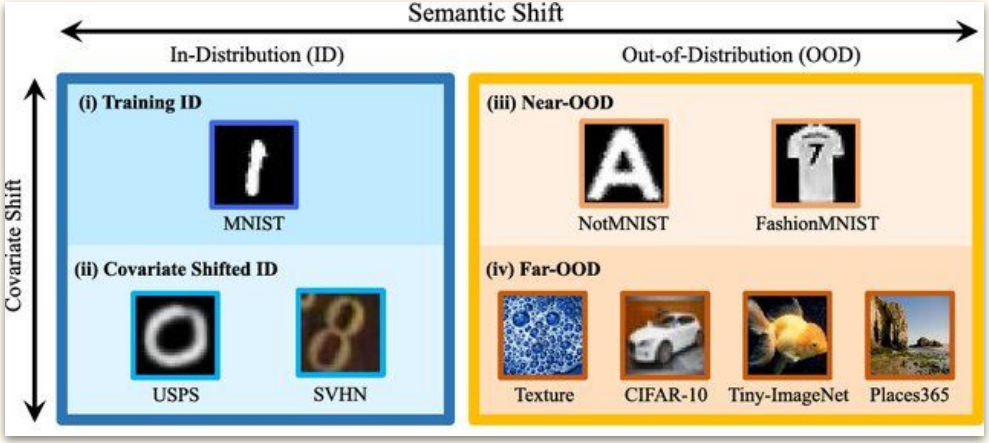


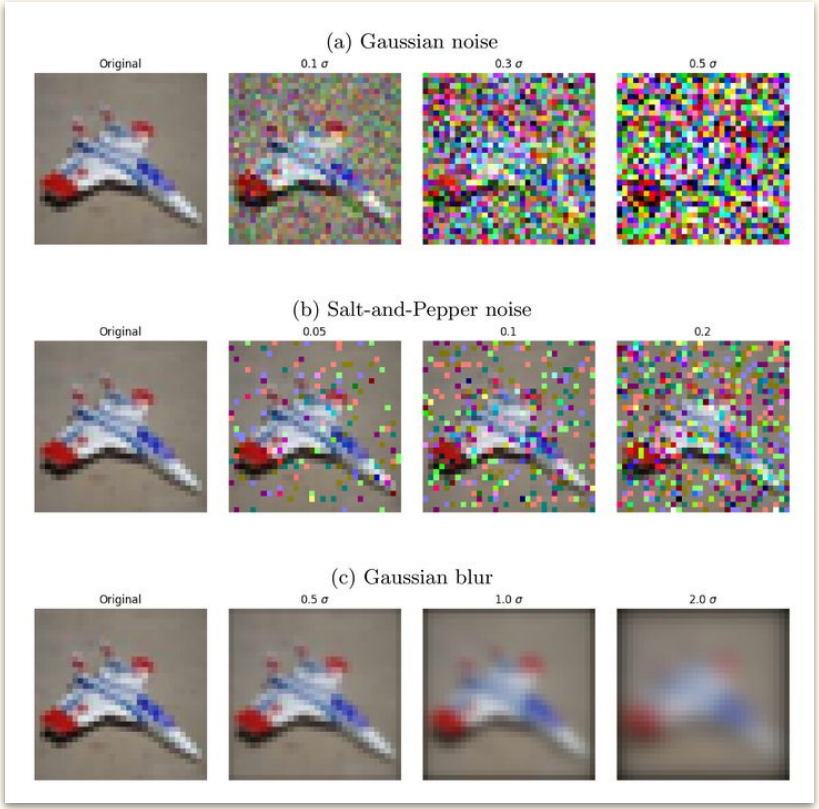
are these really OOD? is it fair to treat them the same as images which are unambiguously OOD?

Labels in OOD detection

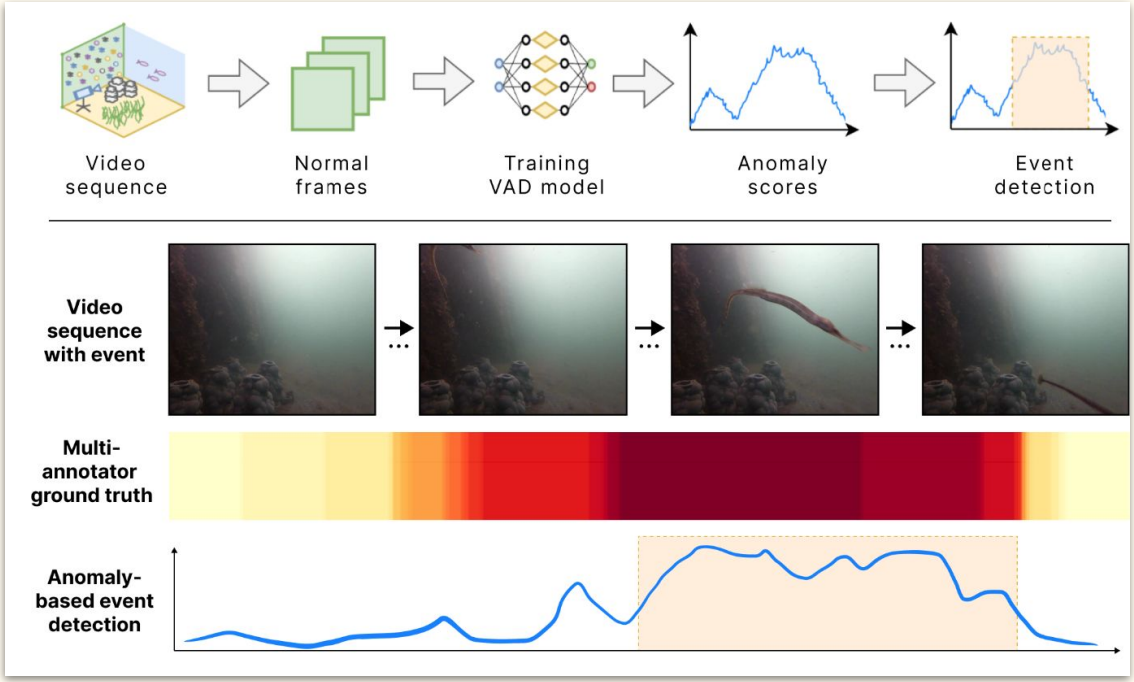
The OOD detection problem is not always as clear-cut in practice as it is on paper...

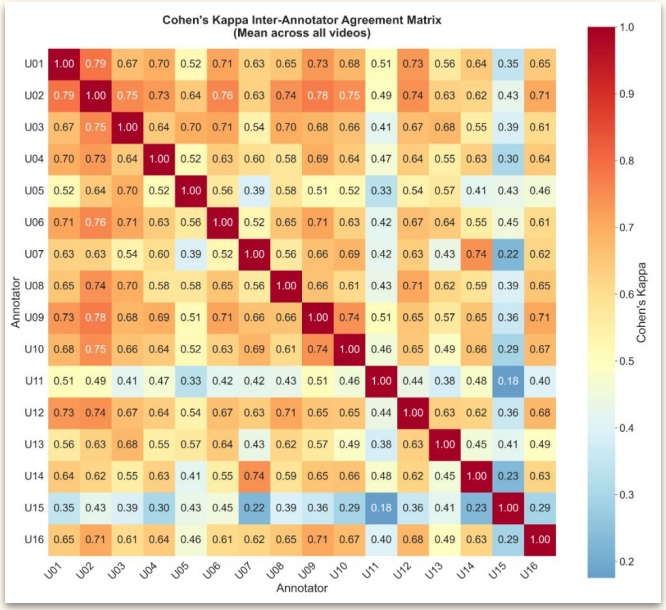
At what point does an ID image become OOD?





Multi-annotator anomaly labels for VAD





[2510.10750] Uncovering Anomalous Events for Marine Environmental Monitoring via Visual Anomaly Detection, ICCVW, 2025.

We should zoom in more on model failures...

...but this requires more fine-grained insights about the data than we typically have. E.g. which samples are more difficult/prone to model errors and why? should we reconsider the original annotations used for evaluation?

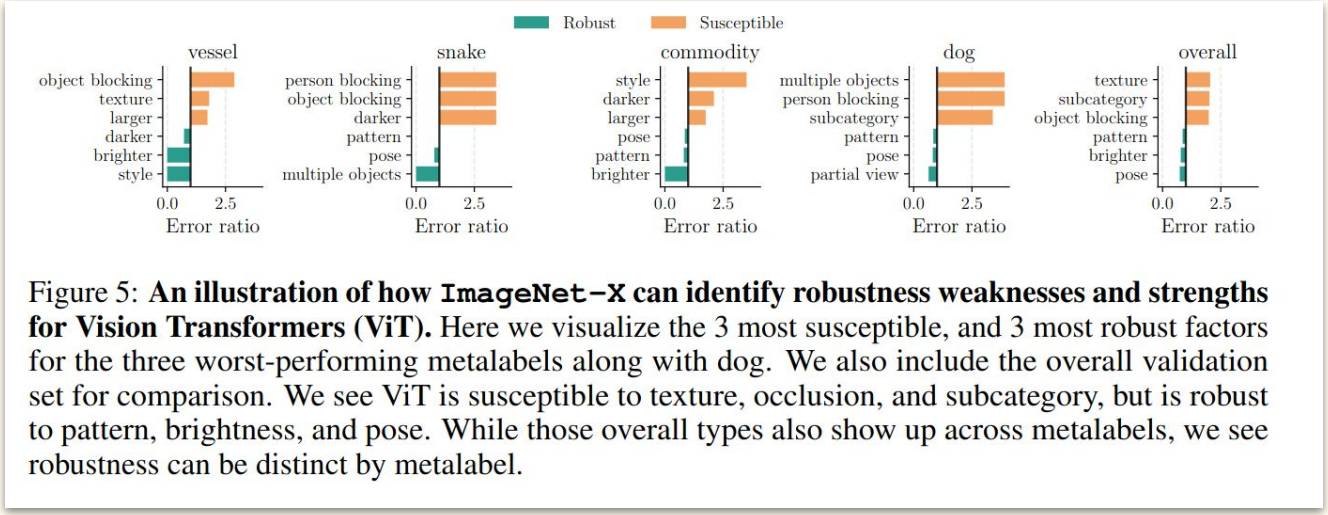


Figure 5: **An illustration of how `ImageNet-X` can identify robustness weaknesses and strengths for Vision Transformers (ViT).** Here we visualize the 3 most susceptible, and 3 most robust factors for the three worst-performing metalabels along with dog. We also include the overall validation set for comparison. We see ViT is susceptible to texture, occlusion, and subcategory, but is robust to pattern, brightness, and pose. While those overall types also show up across metalabels, we see robustness can be distinct by metalabel.

[2211.01866] ImageNet-X: Understanding Model Mistakes with Factor of Variation Annotations ICLR, 2023.

Wrapping up

Beyond image classification

- mainly focused on image-level or pixel-level classification tasks here
- this is also what the noisy label and UQ literature largely focuses on
- increased task complexity/dimensions leads to more complex error & uncertainty profiles
    - we should move beyond CIFAR-style datasets
    - e.g. object detection can have different types of labeling errors: missing bounding box, duplicate bounding box, incorrect class, imprecise localization
- there is a need for uncertainty measures and evaluation criteria which disentangle these

Overall takeaways

- labels should receive more attention.
- we should learn from other fields e.g. NLP, which (IMO) is ahead in terms of discussing & embracing label uncertainty
- when collecting a dataset, consider trying to characterize uncertainty (e.g. multi-annotator labels, comparison to a golden truth) and recording valuable metadata (e.g. annotator background, uncertainty, interactions)
- explore methods which explicitly model aleatoric uncertainty
- don't forget that test labels can also be imperfect/ambiguous! even the ones used to evaluate UQ methods...

Imperfect and uncertain labels in computer vision

Gala

gegeh@create.aau.dk

AALBORG
UNIVERSITY