From ImageNet to Image Classification:

Contextualizing Progress on Benchmarks

Dimitris Tsipras | Shibani Santurkar | Logan Engstrom | Andrew Ilyas | Aleksander Ma¸dry

Review presented by Mounica Subramani

Overview

- Problem Statement
- Original Work
- ImageNet Dataset
  - ImageNet creation pipeline
  - Revisiting ImageNet Labels
- Proposed Work
- Framework – From Label Validation to Image Classification
  - Obtaining Candidate Labels
  - Image Classification via CLASSIFY task
- Quantifying benchmark-task alignment of ImageNet
  - Multi-object images
  - Bias in label validation
- Model Evaluation

Problem Statement

This paper takes a step towards understanding how closely the widely-used vision benchmarks align with real-world tasks, that they are meant to approximate.

Building rich Machine Learning datasets in a scalable manner often necessitates a crowd-sourced data collection pipeline.

This paper analysis pinpoints, how a noisy data collection pipeline can lead to a systematic misalignment between the resulting benchmark and the real-world task it serves as a proxy for.

Original Work

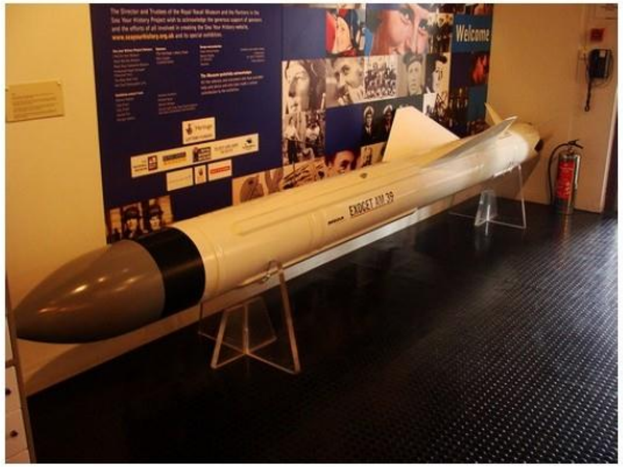
Data created through scalable methods, like building a dataset with the help of large group of people (Annotators)

Annotators were not asked to classify images, but rather to validate a specific automatically-obtained candidate label without knowledge of other classes in the dataset.

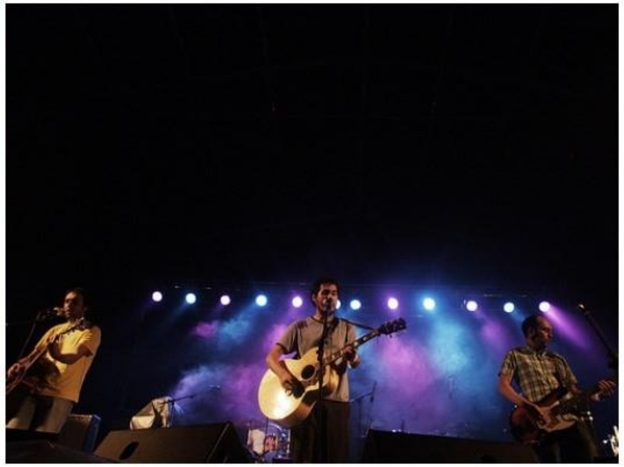
Leads to systemic annotation issues in the dataset, multi-object image problem and bias in the label validation

These issues results in tension between building realistic large-scale benchmarks and the scalable data collection pipeline

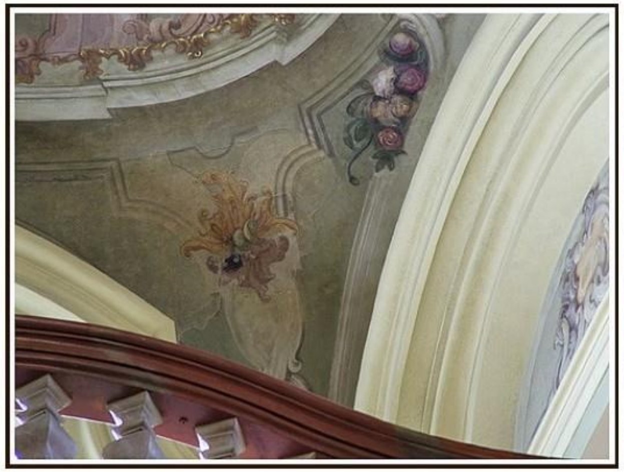
Correct Labels?



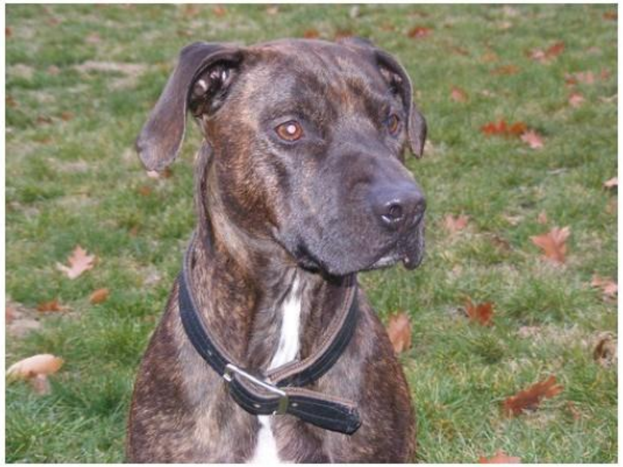
~~(a) missile~~



~~(b) stage~~



~~(c) monastery~~



~~(d) Staffordshire bull terrier~~

ImageNet Labels

projectile

acoustic guitar

church

American Staffordshire terrier

ImageNet Dataset

ImageNet is a prototypical example of large-scale dataset

It has 1000 classes and more than million images

Created through automated data collection and crowd-sourced filtering

Creation Pipeline has two stages

- Image and Label collection
- Image validation via *CONTAINS* task

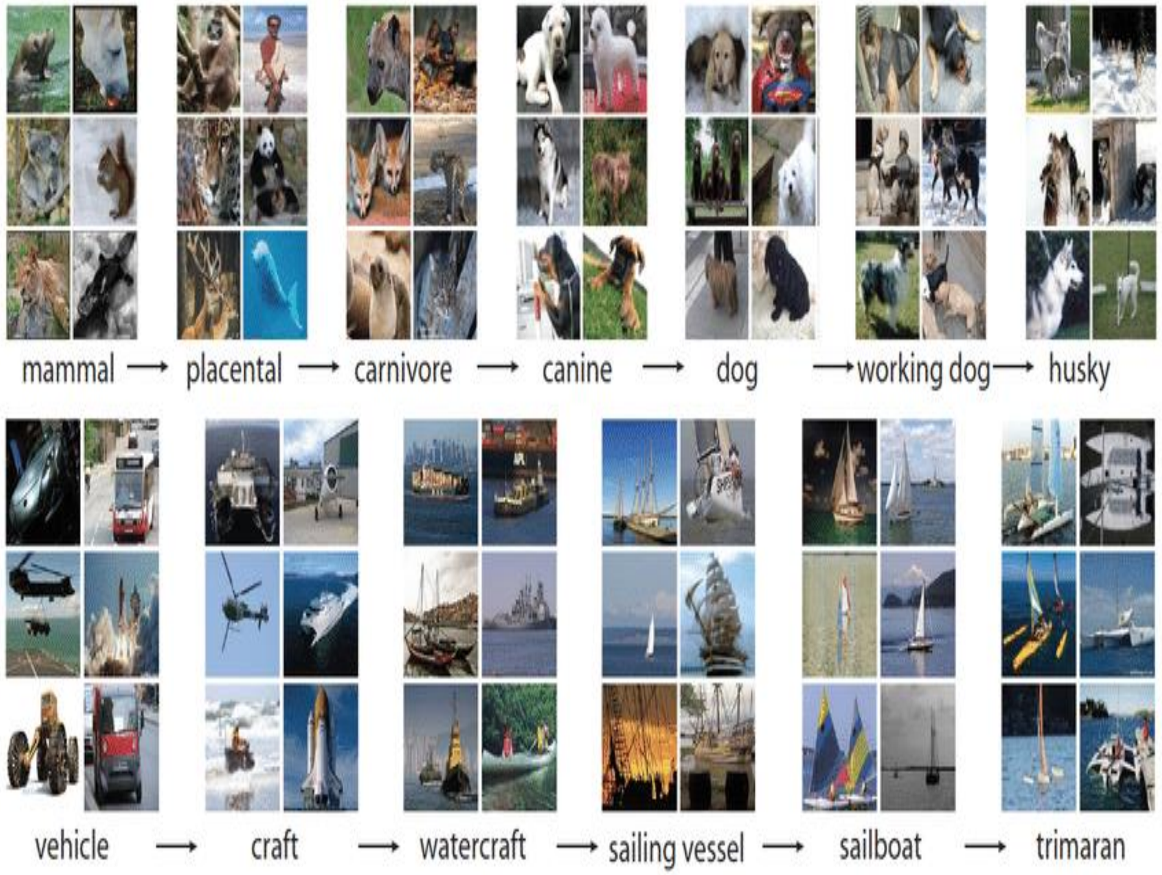

mammal → placental → carnivore → canine → dog → working dog → husky
vehicle → craft → watercraft → sailing vessel → sailboat → trimaran

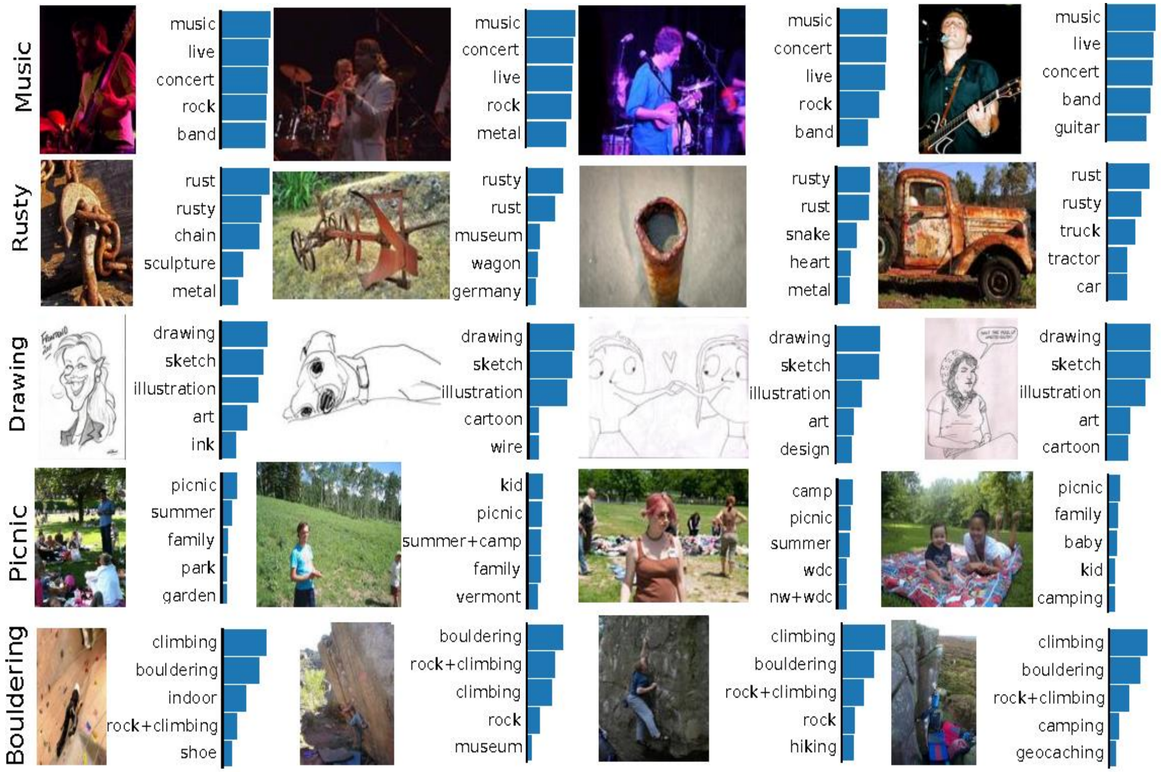
Image and label collection

Selected a set of classes using WordNet hierarchy

For each class, sourced images by querying several search engines

To expand image pool, performed queries in several languages

Label for image is given by the WordNet node that was used for query search



: Single-tag retrieval results, and automatically-generated annotations. None of the query tags are in NUS-WIDE, and most (m

Image Validation – *CONTAINS* task

Employed annotators via the Mechanical Turk (MTurk) crowd-sourcing platform.

Annotators were presented with its class description and a grid of candidate images.

Select all images in the grid that contained an object of that class.

Only the images that received a “convincing majority” of votes were included in the ImageNet.

Annotators were presented with its class description (along with links to the relevant Wikipedia pages) and a grid of candidate images.

Revisiting the ImageNet Labels

Resulting data set might not accurately capture the ground truth.

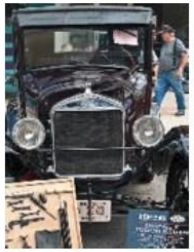
The above discussed pipeline design itself can lead to systematic errors in dataset

The root cause for many of these errors is that the image validation stage (i.e., the *CONTAINS* task) asks annotators only to verify if a specific proposed label is valid for a given image.

Crucially, annotators are never asked to choose among different possible labels for the image and, in fact, have no knowledge of what the other classes even are.

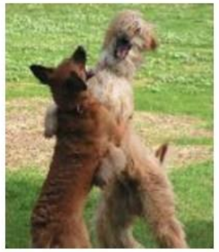
Images with Multiple Objects

- Annotators are instructed to ignore the presence of other objects when validating a particular ImageNet label for an image.



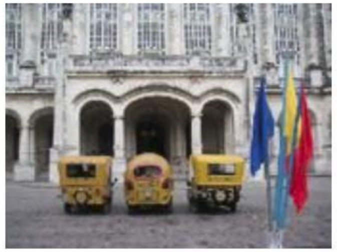
IN label:
Model T

Objects: 3
car wheel
Model T
grille



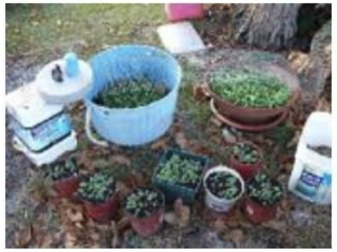
IN label:
Afghan hound

Objects: 2
malinois
Afghan hound



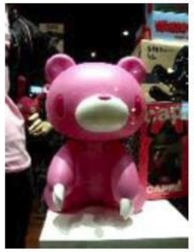
IN label:
cab

Objects: 3
cab
car wheel
palace



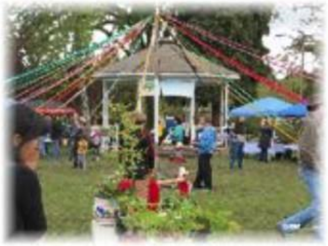
IN label:
pot

Objects: 2
bucket
pot



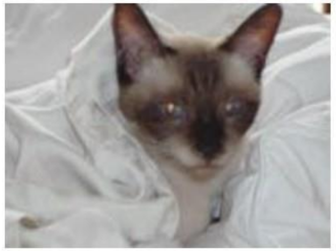
IN label:
toyshop

Objects: 2
teddy
toyshop



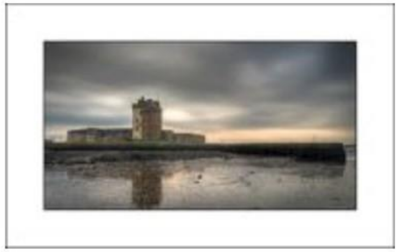
IN label:
maypole

Objects: 2
maypole
pole



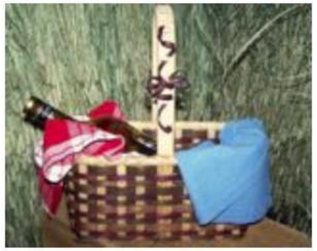
IN label:
Siamese cat

Objects: 2
quilt
Siamese cat



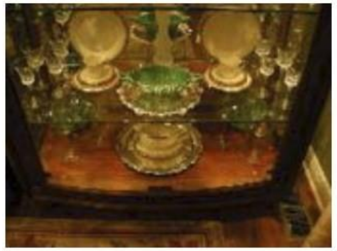
IN label:
castle

Objects: 2
castle
seashore



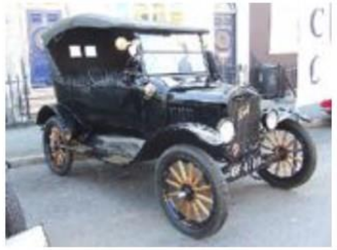
IN label:
hamper

Objects: 2
hamper
wine bottle



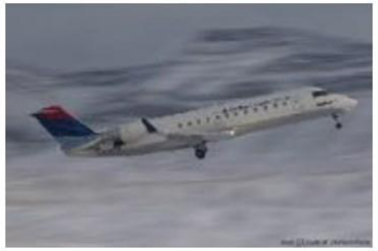
IN label:
china cabinet

Objects: 4
goblet
vase
china cabinet
tray



IN label:
Model T

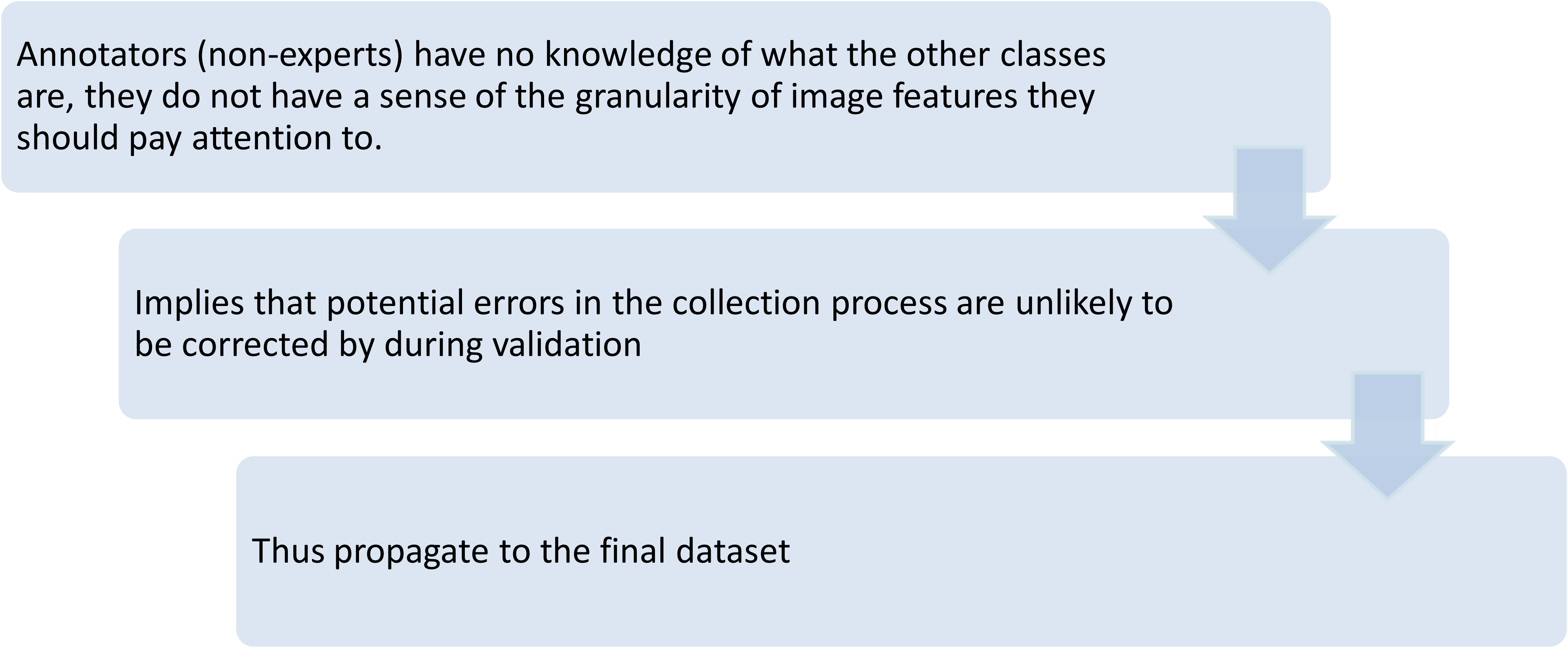
Objects: 4
car wheel
Model T
grille
convertible



IN label:
airliner

Objects: 2
airliner
wing

Biases in Image Filtering



Annotators (non-experts) have no knowledge of what the other classes are, they do not have a sense of the granularity of image features they should pay attention to.

Implies that potential errors in the collection process are unlikely to be corrected by during validation

Thus propagate to the final dataset

Proposed Work

Develop methodology for obtaining fine-grained data annotations via large-scale human studies

Address short fall of object recognition benchmarks

Current standard model evaluation is insufficient

Introducing Human-based performance evaluation

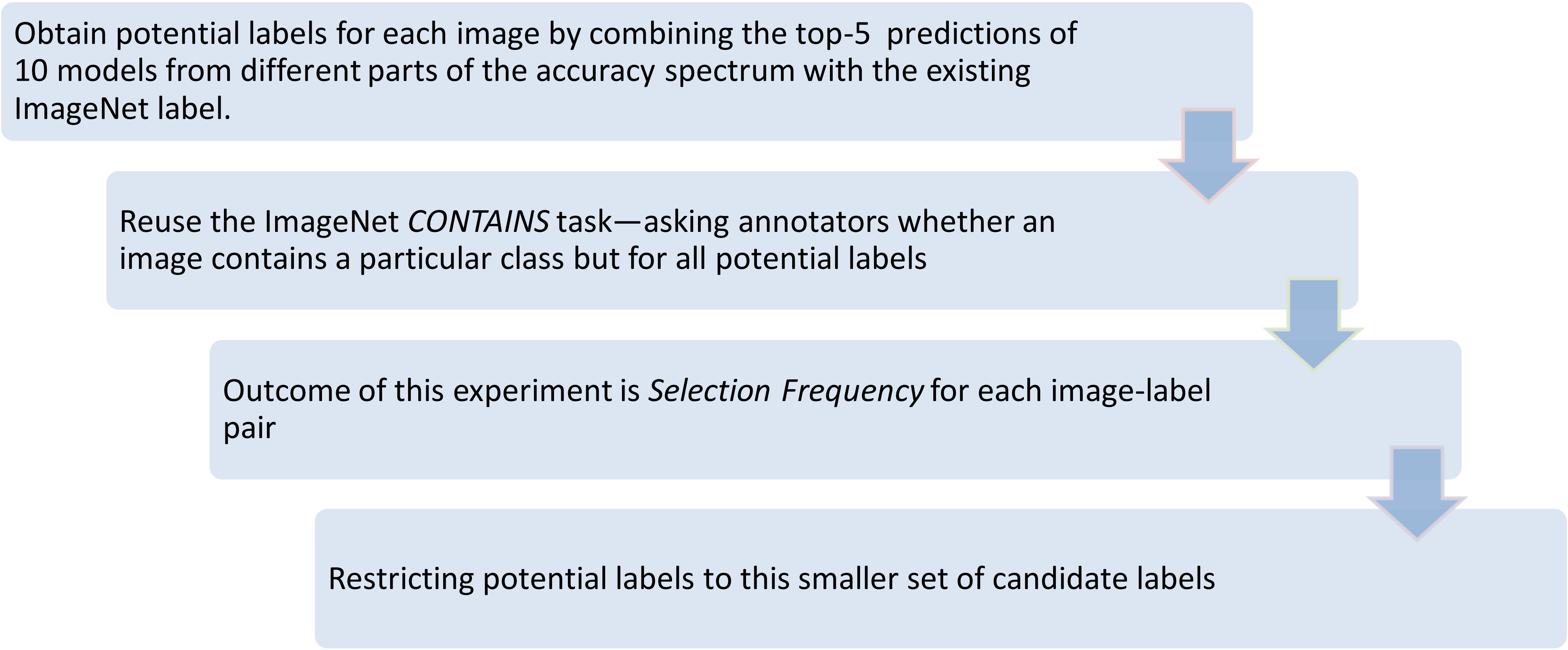
Framework - Label Validation to Image Classification

Authors would like annotators to classify the image, selecting all the relevant labels for it.

There are two phases in this step.

- First, they obtain a small set of (potentially) relevant candidate labels for each image.
- Second, they present these labels to annotators and ask them to select one of them for each distinct object using what they call the *CLASSIFY* task.

Obtaining Candidate Labels



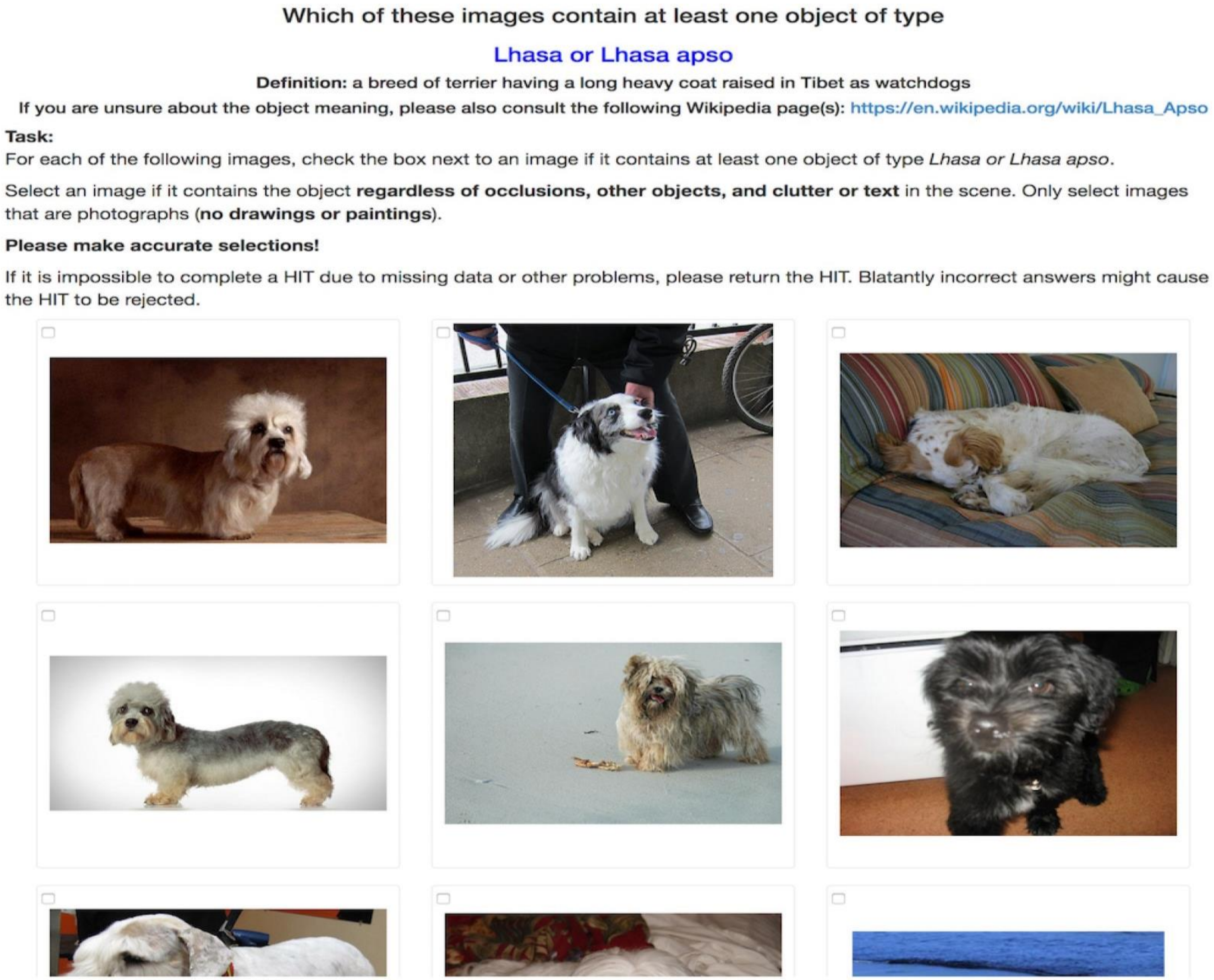
Obtain potential labels for each image by combining the top-5 predictions of 10 models from different parts of the accuracy spectrum with the existing ImageNet label.

Reuse the ImageNet *CONTAINS* task—asking annotators whether an image contains a particular class but for all potential labels

Outcome of this experiment is *Selection Frequency* for each image-label pair

Restricting potential labels to this smaller set of candidate labels

Obtaining Candidate Labels



Which of these images contain at least one object of type

Lhasa or Lhasa apso

Definition: a breed of terrier having a long heavy coat raised in Tibet as watchdogs

If you are unsure about the object meaning, please also consult the following Wikipedia page(s): https://en.wikipedia.org/wiki/Lhasa_Apso

Task:

For each of the following images, check the box next to an image if it contains at least one object of type *Lhasa or Lhasa apso*.

Select an image if it contains the object **regardless of occlusions, other objects, and clutter or text** in the scene. Only select images that are photographs (**no drawings or paintings**).

Please make accurate selections!

If it is impossible to complete a HIT due to missing data or other problems, please return the HIT. Blatantly incorrect answers might cause the HIT to be rejected.

□ □ □

□ □ □

□ □ □

Image Classification via the CLASSIFY Task

Asking annotators to identify

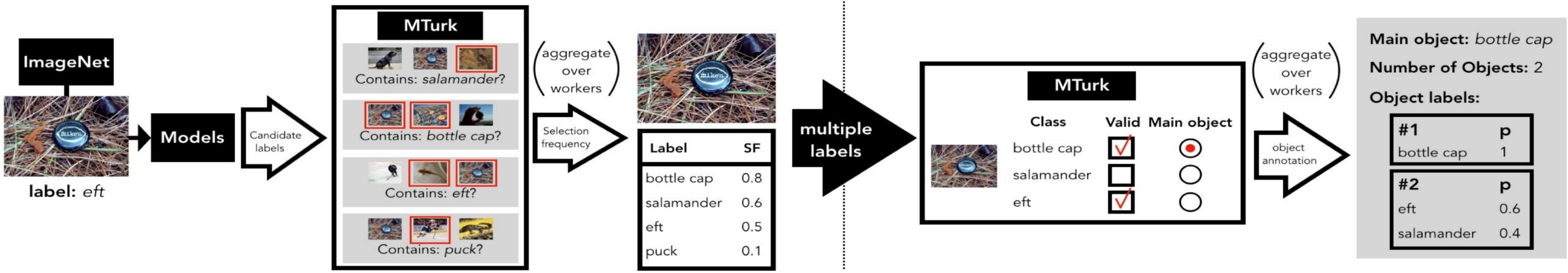
All labels that correspond to objects in the image.

The label for the main object (according to their judgment).

Select only one label per distinct object except mutually exclusive objects.

Identifying the main label and number of objects

From each annotator’s response, authors learn what they consider to be the label of the main object, as well as how many objects they think are present in the image



Label	SF
bottle cap	0.8
salamander	0.6
eft	0.5
puck	0.1

Class	Valid	Main object
bottle cap	☑	●
salamander	□	○
eft	☑	○

Main object: *bottle cap*

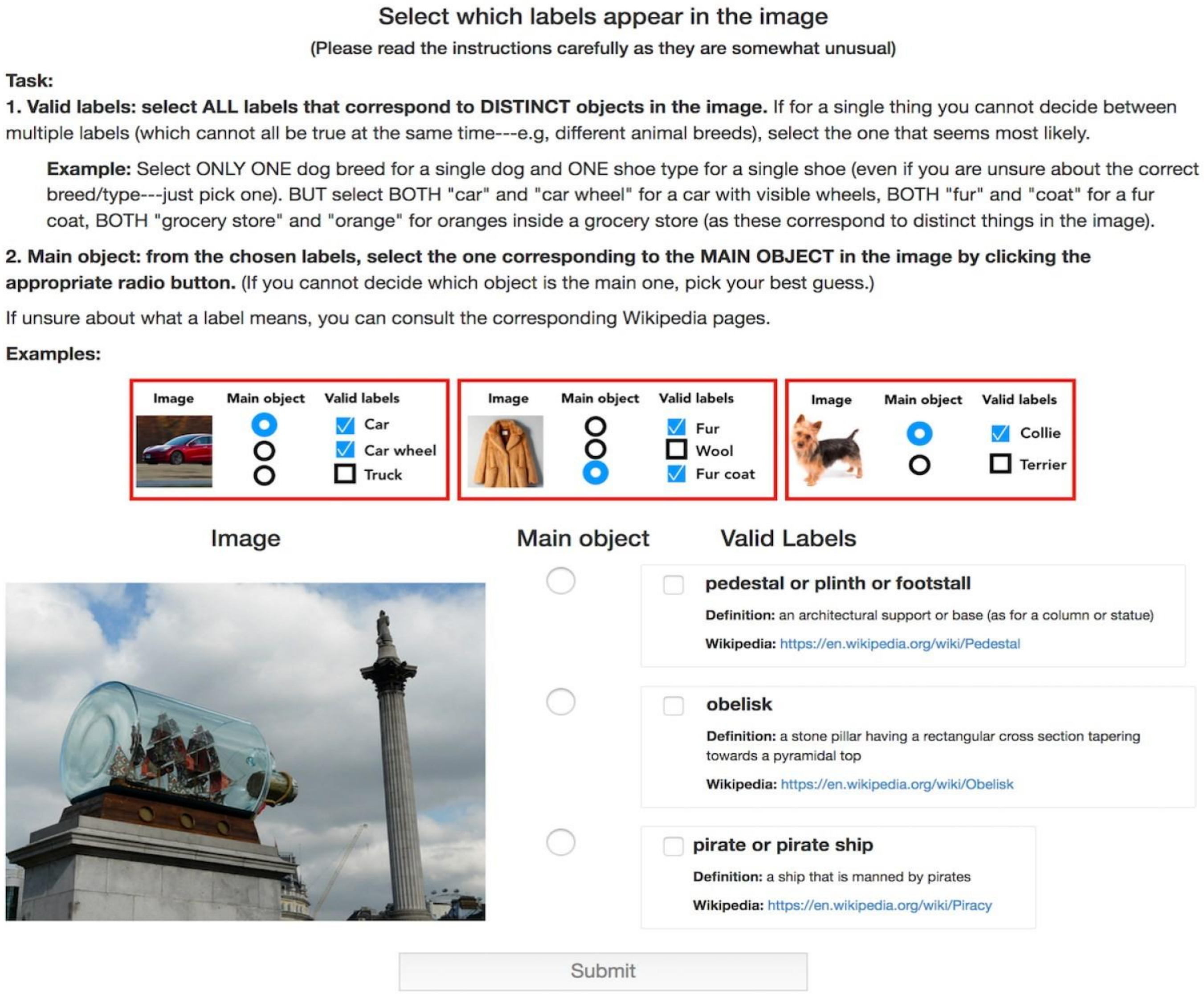
Number of Objects: 2

Object labels:

#1	p
bottle cap	1

#2	p
eft	0.6
salamander	0.4

Image Classification via the CLASSIFY Task



Select which labels appear in the image

(Please read the instructions carefully as they are somewhat unusual)

Task:

1. Valid labels: select ALL labels that correspond to DISTINCT objects in the image. If for a single thing you cannot decide between multiple labels (which cannot all be true at the same time---e.g, different animal breeds), select the one that seems most likely.

> **Example:** Select ONLY ONE dog breed for a single dog and ONE shoe type for a single shoe (even if you are unsure about the correct breed/type---just pick one). BUT select BOTH "car" and "car wheel" for a car with visible wheels, BOTH "fur" and "coat" for a fur coat, BOTH "grocery store" and "orange" for oranges inside a grocery store (as these correspond to distinct things in the image).

2. Main object: from the chosen labels, select the one corresponding to the MAIN OBJECT in the image by clicking the appropriate radio button. (If you cannot decide which object is the main one, pick your best guess.)

If unsure about what a label means, you can consult the corresponding Wikipedia pages.

Examples:

Image	Main object	Valid labels
	●	☑ Car
	○	☑ Car wheel
	○	□ Truck

Image	Main object	Valid labels
	○	☑ Fur
	○	□ Wool
	●	☑ Fur coat

Image	Main object	Valid labels
	●	☑ Collie
	○	□ Terrier

Image	Main object	Valid Labels
	○	□ **pedestal or plinth or footstall** **Definition:** an architectural support or base (as for a column or statue) **Wikipedia:** https://en.wikipedia.org/wiki/Pedestal
	○	□ **obelisk** **Definition:** a stone pillar having a rectangular cross section tapering towards a pyramidal top **Wikipedia:** https://en.wikipedia.org/wiki/Obelisk
	○	□ **pirate or pirate ship** **Definition:** a ship that is manned by pirates **Wikipedia:** https://en.wikipedia.org/wiki/Piracy

Quantifying the Benchmark-Task Alignment of ImageNet

- Multi-object images
    - More than 1/5 of the ImageNet dataset contains at least two objects.
    - Models perform significantly worse on multi-label images based on top-1 accuracy (measured w.r.t. ImageNet labels): accuracy drops by more than 10% across all models.
    - There are pairs of classes which consistently co-occur.
    - Model accuracy is especially low on certain classes that systematically co- occur.
    - A more natural notion of accuracy for multi-object images would be, to consider a model prediction to be correct if it matches the label of any object in the image.

Multi-Object Images

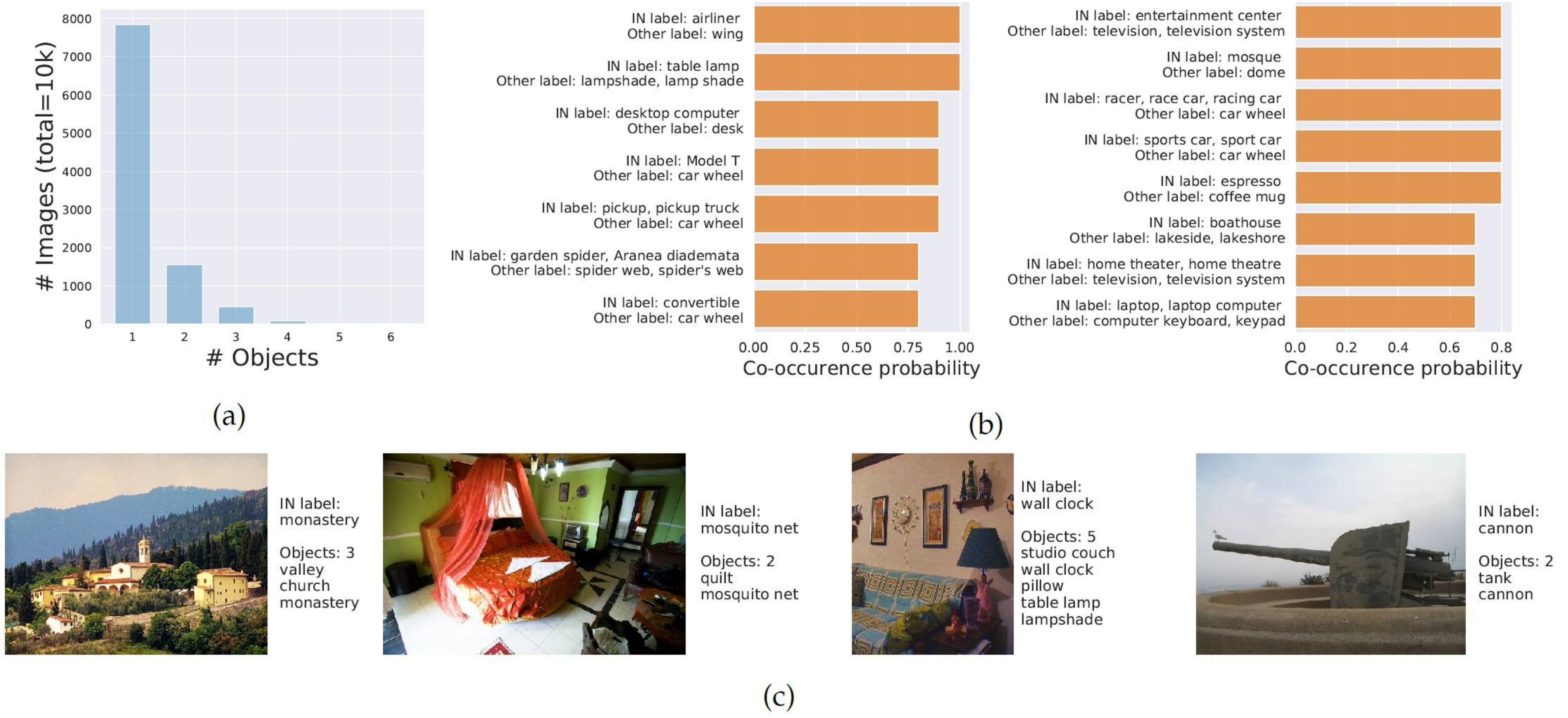


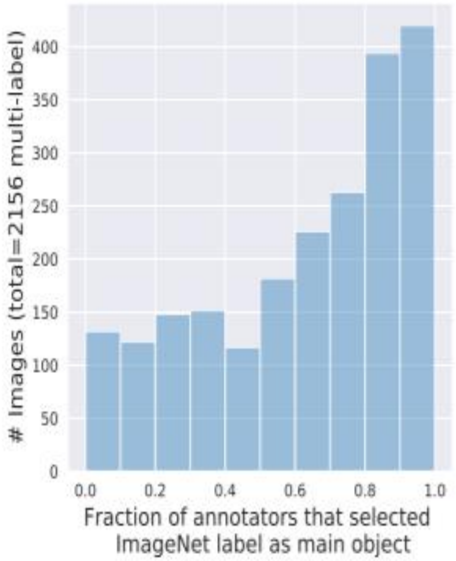
Figure 3: (a) Number of objects per image—more than a fifth of the images contains two or more objects from ImageNet classes. (b) Pairs of classes which consistently co-occur as distinct objects. Here, we visualize the top 15 ImageNet classes based on how often their images contain another *fixed* object ("Other label"). (c) Random examples of multi-label ImageNet images (cf. Appendix Figure 17 for additional samples).

Multi-Object Images: Human-label disagreement

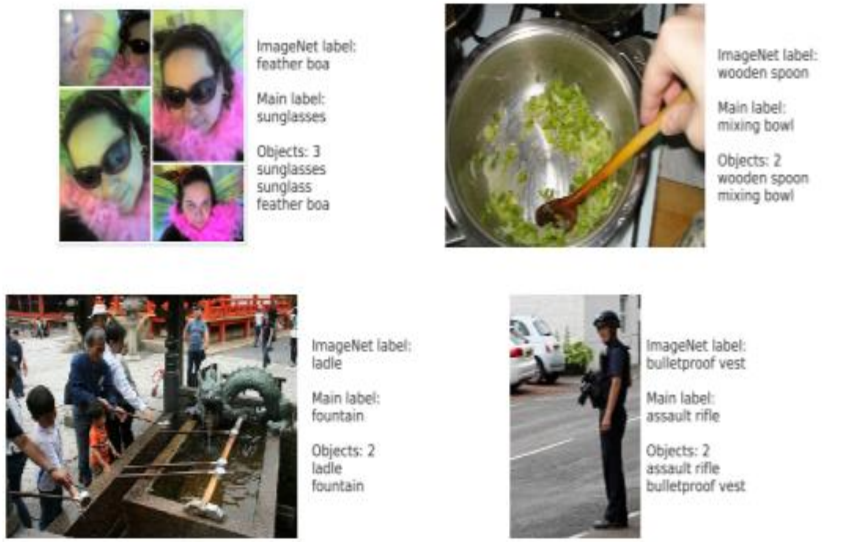
Although models suffer a sizeable accuracy drop on multi-object images, they are still relatively good at predicting the ImageNet label.

This bias could be justified whenever there is a distinct main object in the image, and it corresponds to the ImageNet label.

For nearly a third of the multi-object images, the ImageNet label does not denote the most likely main object as judged by human annotators.



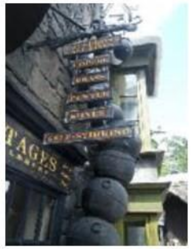
(a)



ImageNet label: feather boa
Main label: sunglasses
Objects: 3 sunglasses sunglass feather boa
ImageNet label: wooden spoon
Main label: mixing bowl
Objects: 2 wooden spoon mixing bowl
ImageNet label: ladle
Main label: fountain
Objects: 2 ladle fountain
ImageNet label: bulletproof vest
Main label: assault rifle
Objects: 2 assault rifle bulletproof vest

(b)

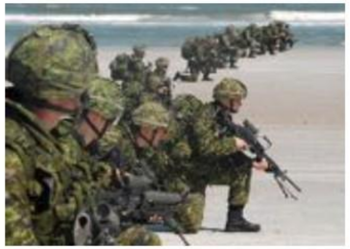
Human-Label disagreement



IN label:
caldron

Main label:
street sign

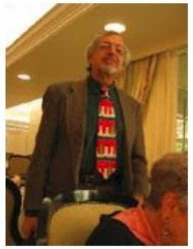
Objects: 2
street sign
caldron



IN label:
assault rifle

Main label:
military uniform

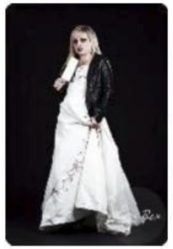
Objects: 3
assault rifle
military uniform
rifle



IN label:
Windsor tie

Main label:
suit

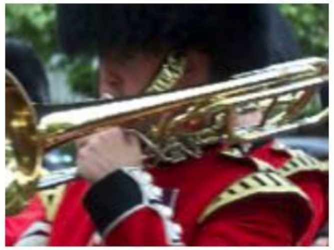
Objects: 2
Windsor tie
suit



IN label:
cleaver

Main label:
gown

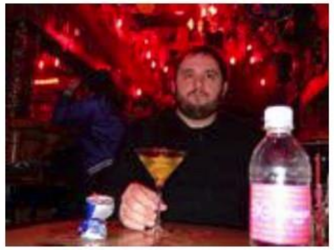
Objects: 2
cleaver
gown



IN label:
bearskin

Main label:
trombone

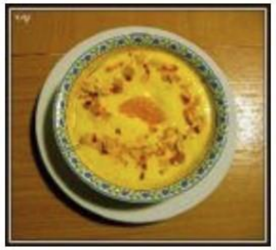
Objects: 2
bearskin
trombone



IN label:
water bottle

Main label:
restaurant

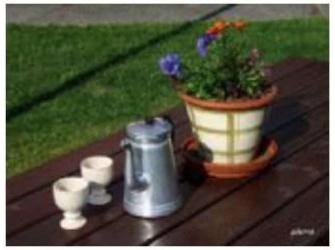
Objects: 3
goblet
restaurant
water bottle



IN label:
plate

Main label:
soup bowl

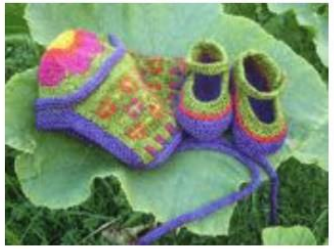
Objects: 2
plate
soup bowl



IN label:
coffeepot

Main label:
teapot

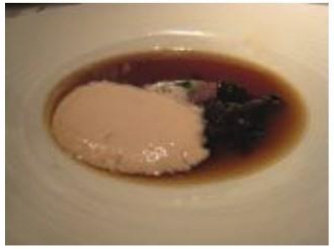
Objects: 3
teapot
cup
pot



IN label:
bonnet

Main label:
wool

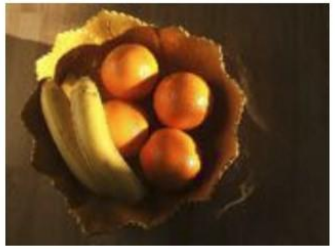
Objects: 2
wool
bonnet



IN label:
consomme

Main label:
plate

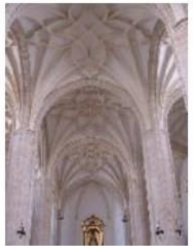
Objects: 2
consomme
soup bowl



IN label:
banana

Main label:
orange

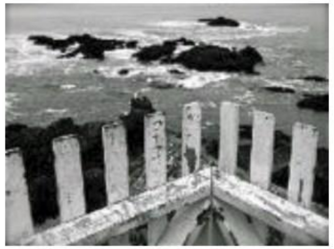
Objects: 2
orange
banana



IN label:
vault

Main label:
church

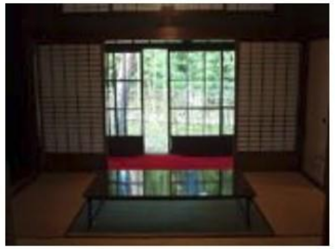
Objects: 2
altar
church



IN label:
picket fence

Main label:
seashore

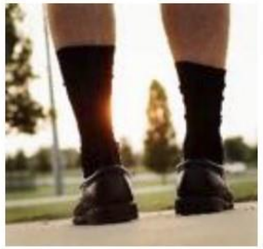
Objects: 2
picket fence
seashore



IN label:
shoji

Main label:
sliding door

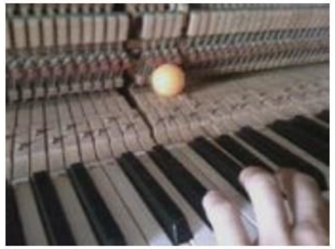
Objects: 4
sliding door
shoji
dining table
window shade



IN label:
sock

Main label:
Loafer

Objects: 2
sock
Loafer



IN label:
ping-pong ball

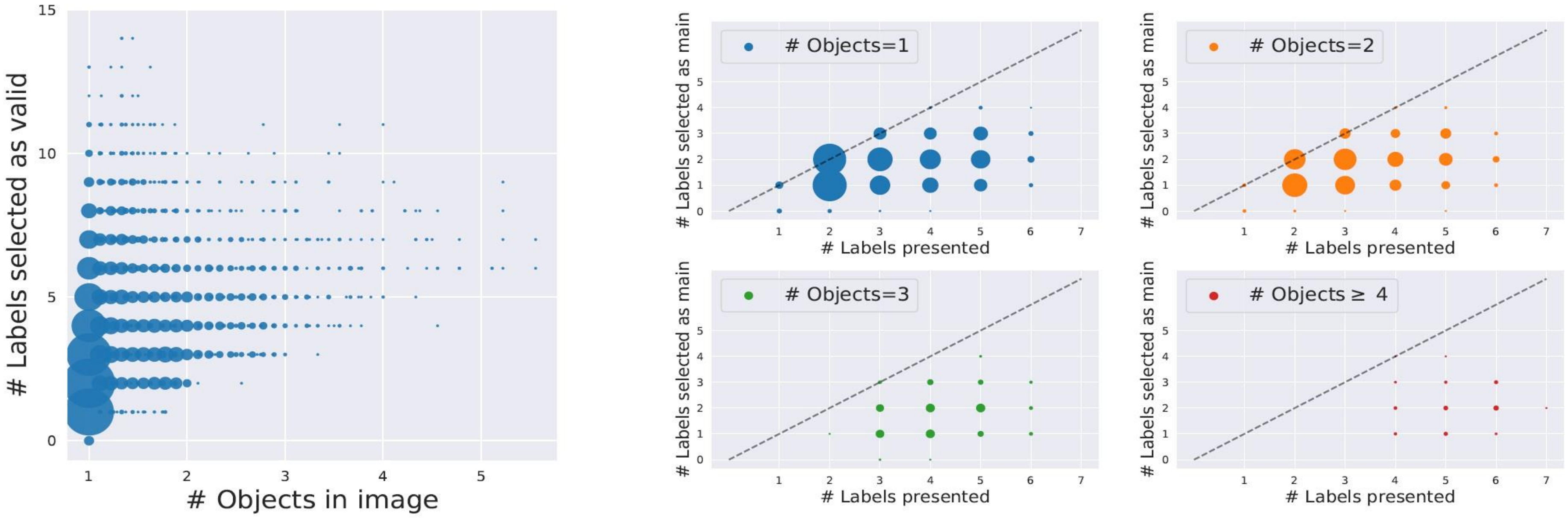
Main label:
upright

Objects: 2
ping-pong ball
upright

Quantifying the Benchmark-Task Alignment of ImageNet

- Bias in label validation
  - Recall that annotators are asked a somewhat leading question, of whether a specific label is present in the image, making them prone to answering positively even for images of a different, yet similar, class.
  - Under the original task setup (i.e., the CONTAINS task), annotators consistently select multiple labels, for nearly 40% of the images, another label is selected at least as often as the ImageNet label.
  - Even when annotators perceive a single object, they often select as many as 10 classes.
  - If instead of asking annotators to judge the validity of a specific label(s) in isolation, asking them to choose the main object among several possible labels simultaneously (i.e., via the CLASSIFY task), they select substantially fewer labels.

Bias in Label Validation



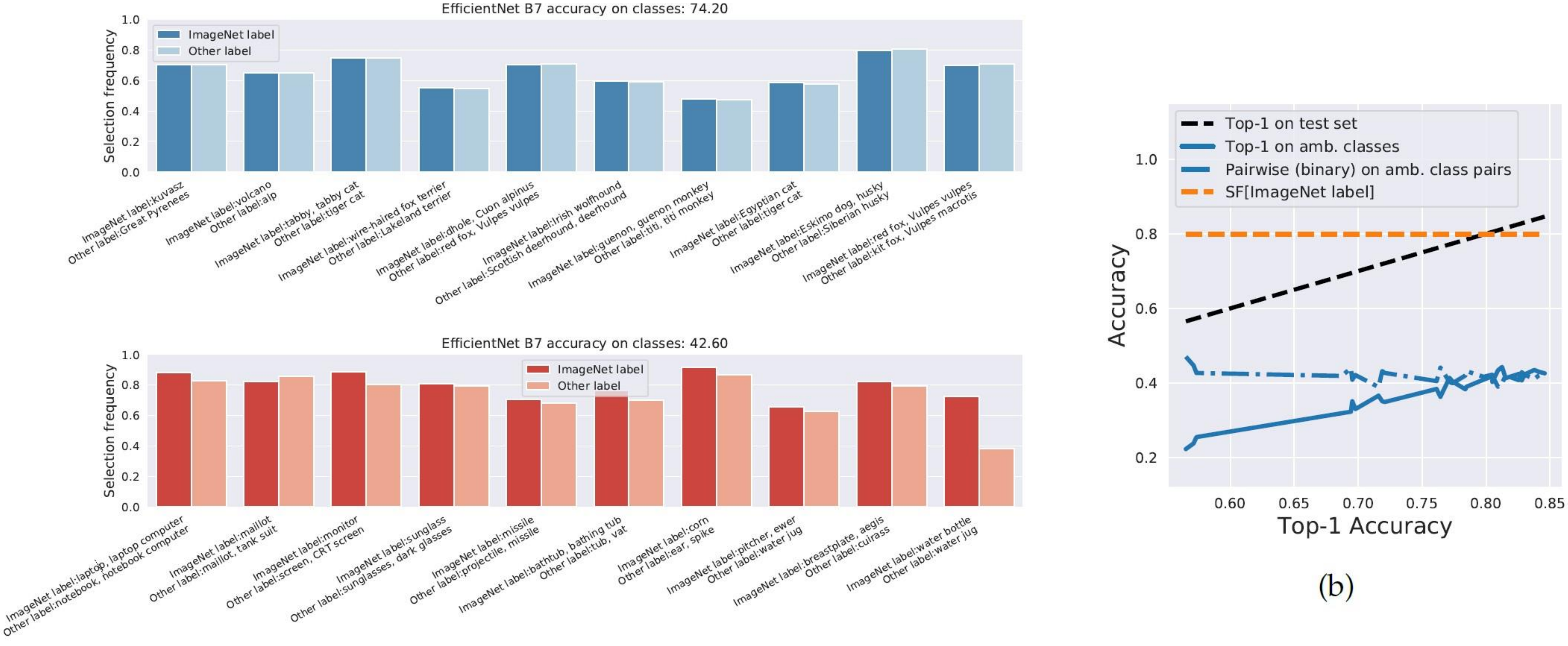
(a)

(b)

Figure 8: (a) Number of labels selected in the CONTAINS task: the y-axis measures the number of labels that were selected by at least two annotators; the x-axis measures the average number of objects indicated to be present in the image during the CLASSIFY task; the dot size represents the number of images in each 2D bin. Even when annotators perceive an images as depicting a single object, they often select multiple labels as valid. (b) Number of labels that at least two annotators selected for the main image object (in the CLASSIFY task; cf. Section 3.2) as a function of the number of labels presented to them. Annotator confusion decreases significantly when the task setup explicitly involves choosing between multiple labels *simultaneously*.

Bias in Label Validation

- Confusing class pairs
  - There are several pairs of ImageNet classes that annotators have trouble telling apart—they consistently select both labels as valid for images of either class.
  - On some of these pairs we see that models still perform well.
  - However, for other pairs, even state-of-the-art models have poor accuracy (below 40%).
  - If that is indeed the case, it is natural to wonder whether accuracy on ambiguous classes can be improved without overfitting to the ImageNet test set.
  - The annotators' inability to remedy overlaps in the case of ambiguous class pairs could be due to the presence of classes that are semantically quite similar, e.g., "rifle" and "assault rifle".
  - In some cases, there also were errors in the task setup. For instance, there were occasional overlaps in the class names (e.g., "maillot" and "maillot, tank suit") and Wikipedia links (e.g., "laptop computer" and "notebook computer") presented to the annotators.



(a)

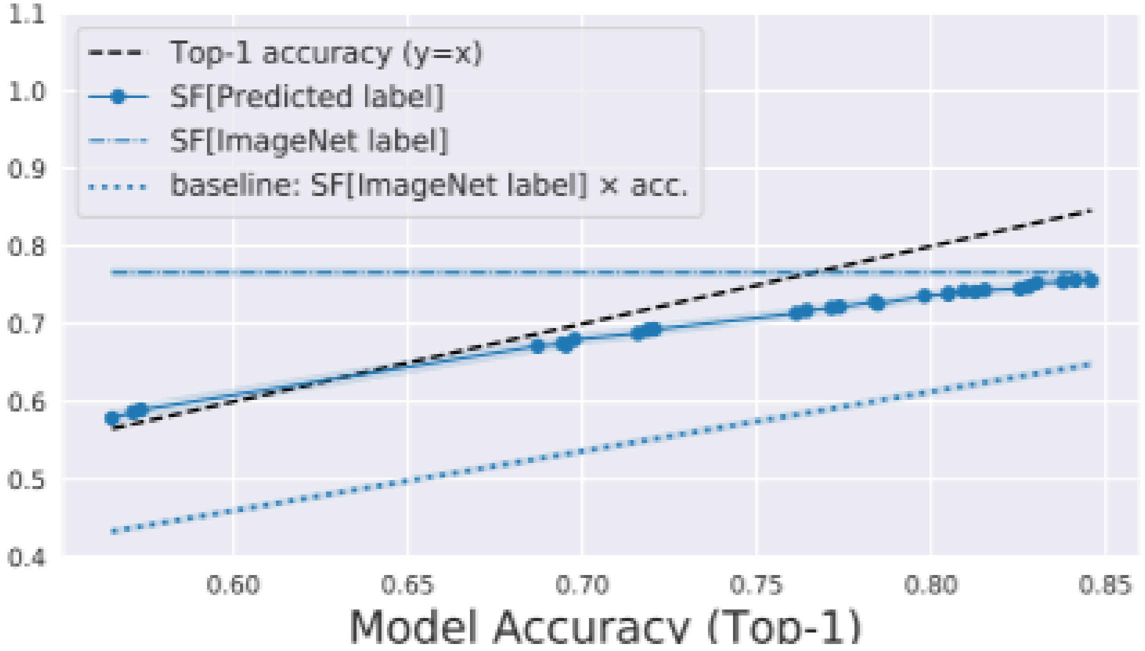
(b)

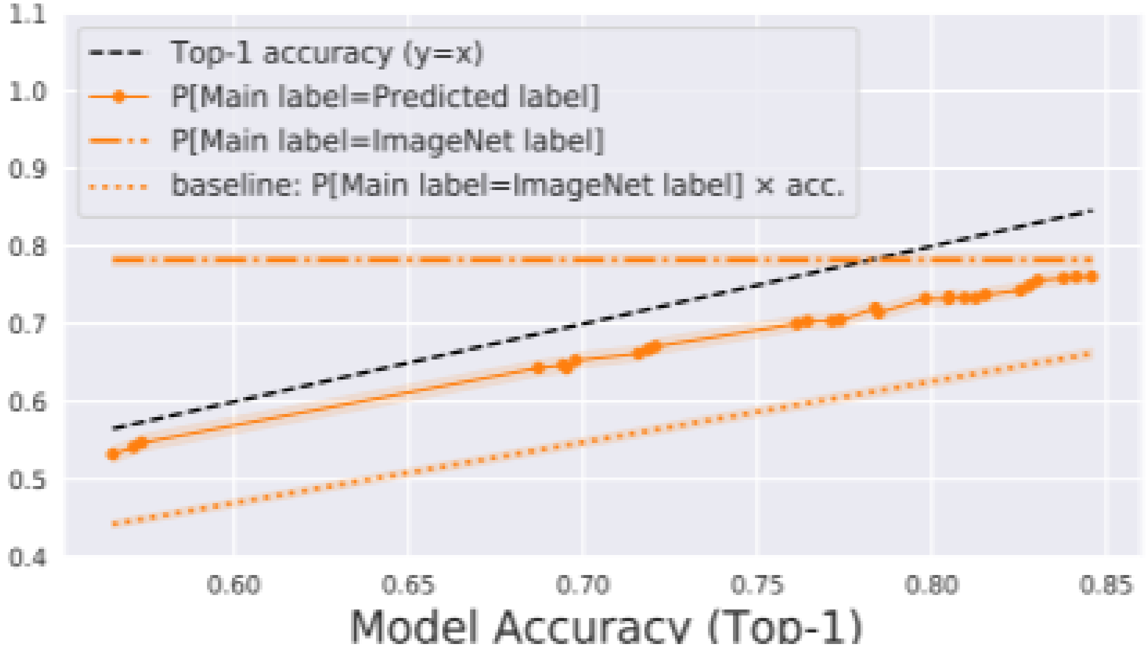
Figure 9: (a) ImageNet class pairs for which annotators often deem both classes as valid. We visualize the top 10 pairs split based on the accuracy of EfficientNet B7 on these pairs being high (*top*) or low (*bottom*). (b) Model progress on ambiguous class pairs (from (a) *bottom*) has been largely stagnant—possibly due to substantial overlap in the class distributions. In fact, models are unable to distinguish between these pairs better than chance (cf. pairwise accuracy).

Model Evaluation - Beyond Test Accuracy

Human assessment of model predictions

- Selection frequency of the prediction
    - How often annotators select the predicted label as being present in the image (determined using the *CONTAINS* task).
- Accuracy based on main label annotation
    - How frequently the prediction matches the main label for the image, as obtained using our *CLASSIFY* task.





Human Assessment of Model Predictions – Incorrect Predictions

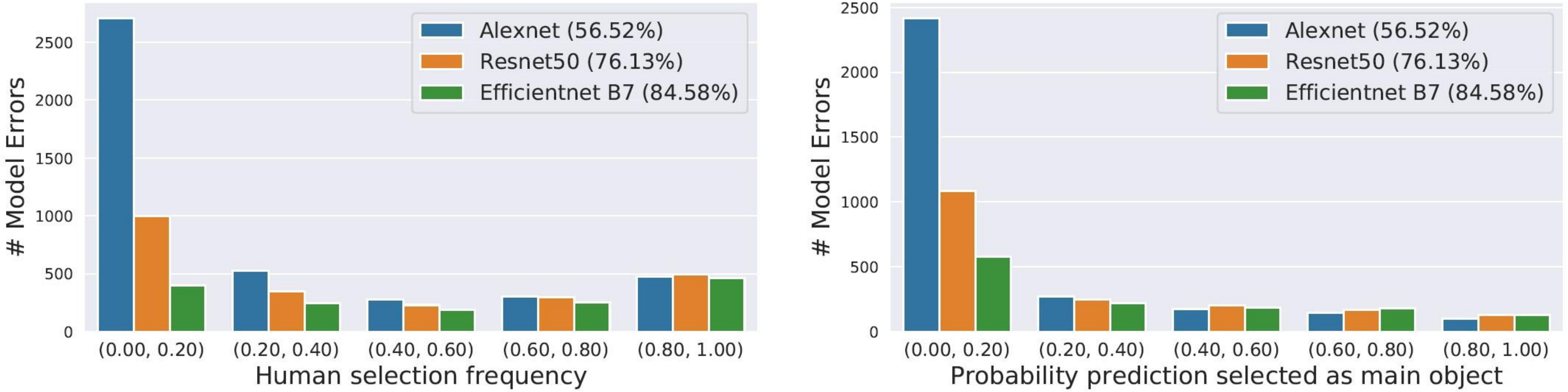


Figure 11: Distribution of annotator selection frequencies (cf. Section 3.1) for model predictions deemed incorrect w.r.t. the ImageNet label. Models that are more accurate also seem to make fewer mistakes that have low human selection frequency (for the corresponding image-prediction pair).

Human Assessment of Model Predictions

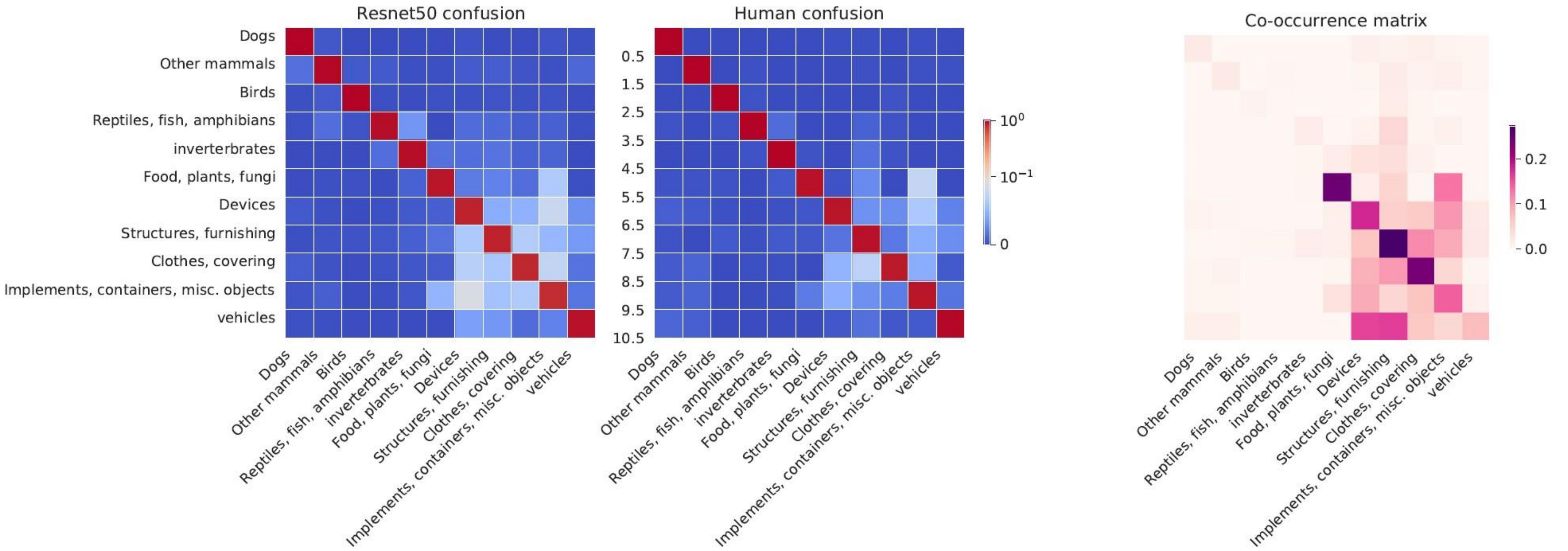
The predictions of state-of-the-art models have gotten, on average, quite close to ImageNet labels.

That is, annotators are almost equally likely to select the predicted label as being present in the image (or being the main object) as the ImageNet label.

This indicates that model predictions might be closer to what non- expert annotators can recognize as the ground truth.

This also means, we are at a point where we may no longer by able to easily identify (e.g., using crowd-sourcing) the extent to which further gains in accuracy correspond to such improvements.

Confusion Matrix



(a) Confusion matrices

(b) Co-occurrence matrix

Figure 12: Characterizing similarity between model and human predictions (indicated by main object selection in the CONTAINS task): (a) Model (ResNet-50) and human confusion matrices on 11 ImageNet superclasses (cf. Section A.1.1). (b) Superclass co-occurrence matrix: how likely a specific pair of superclasses is to occur together (based on annotation from Section 3.2).

Conclusion

- There are many images with multiple valid labels, and ambiguous classes in ImageNet.
- Top-1 accuracy often underestimates the performance of models by unduly penalizing them for predicting the label of a different, but also valid, image object.
- Taking a step towards evaluation metrics that overcome these issues, authors utilize human annotators to directly judge the correctness of model predictions.
- On the positive side, they find that models that are more accurate on ImageNet also tend to be more human-aligned in their errors.
- While this might be reassuring, it also indicates that we are at a point where we cannot easily gauge (e.g., via simple crowd-sourcing) whether further progress on the ImageNet benchmark is meaningful, or is simply a result of overfitting to this benchmarks' particular characterictis.