

## A. Experimental setup

### A.1. Datasets

We perform our analysis on the ImageNet dataset (Russakovsky et al., 2015). A full description of the data creation process can be found in Deng et al. (2009) and Russakovsky et al. (2015). For the purposes of our human studies, we use a random subset of the validation set—10,000 images chosen by sampling 10 random images from each of the 1,000 classes. (Model performance on this subset closely mirrors overall test accuracy, as shown in Figure 4.) In our analysis, we refer to the original dataset labels as “ImageNet labels” or “IN labels”.

#### A.1.1. IMAGENET SUPERCLASSES

We construct 11 superclasses by manually grouping together semantically similar 1000 ImageNet classes (in parenthesis), with the assistance of the WordNet hierarchy—*Dogs* (130), *Other mammals* (88), *Bird* (59), *Reptiles, fish, amphibians* (60), *Invertebrates* (61), *Food, plants, fungi* (63), *Devices* (172), *Structures, furnishing* (90), *Clothes, covering* (92), *Implements, containers, misc. objects* (117), *Vehicles* (68).

### A.2. Models

We perform our evaluation on various standard ImageNet-trained models—see Appendix Table 1 for a full list. We use open-source pre-trained implementations from [github.com/Cadene/pretrained-models.pytorch](https://github.com/Cadene/pretrained-models.pytorch) and/or [github.com/rwightman/pytorch-image-models/tree/master/timm](https://github.com/rwightman/pytorch-image-models/tree/master/timm) for all architectures.

Table 1: Models used in our analysis with the corresponding ImageNet top-1/5 accuracies.

Model	Top-1	Top-5
alexnet (Krizhevsky et al., 2012)	56.52	79.07
squeezeNet1_1 (Iandola et al., 2016)	57.12	80.13
squeezeNet1_0 (Iandola et al., 2016)	57.35	79.89
vgg11 (Simonyan & Zisserman, 2015)	68.72	88.66
vgg13 (Simonyan & Zisserman, 2015)	69.43	89.03
inception_v3 (Szegedy et al., 2016)	69.54	88.65
googlenet (Szegedy et al., 2015)	69.78	89.53
vgg16 (Simonyan & Zisserman, 2015)	71.59	90.38
mobilenet_v2 (Sandler et al., 2018)	71.88	90.29
vgg19 (Simonyan & Zisserman, 2015)	72.07	90.74
resnet50 (He et al., 2016)	76.13	92.86
efficientnet_b0 (Tan & Le, 2019)	76.43	93.05
densenet161 (Huang et al., 2017)	77.14	93.56
resnet101 (He et al., 2016)	77.37	93.55
efficientnet_b1 (Tan & Le, 2019)	78.38	94.04
wide_resnet50_2 (Zagoruyko & Komodakis, 2016)	78.47	94.09
efficientnet_b2 (Tan & Le, 2019)	79.81	94.73
gluon_resnet152_v1d (He et al., 2019)	80.49	95.17
inceptionresnetv2 (Szegedy et al., 2017)	80.49	95.27
gluon_resnet152_v1s (He et al., 2019)	80.93	95.31
senet154 (Hu et al., 2018)	81.25	95.30
efficientnet_b3 (Tan & Le, 2019)	81.53	95.65
nasnetalarge (Zoph et al., 2018)	82.54	96.01
pnasnet5large (Liu et al., 2018)	82.79	96.16
efficientnet_b4 (Tan & Le, 2019)	83.03	96.34
efficientnet_b5 (Tan & Le, 2019)	83.78	96.71
efficientnet_b6 (Tan & Le, 2019)	84.13	96.96
efficientnet_b7 (Tan & Le, 2019)	84.58	97.00

## B. Obtaining Image Annotations

Our goal is to use human annotators to obtain labels for each distinct object in ImageNet images (provided it corresponds to a valid ImageNet class). To make this classification task feasible, we first identify a small set of relevant candidate labels per image to present to annotators.

### B.1. Obtaining candidate labels

As discussed in Section 3.1, we narrow down the candidate labels for each image by (1) restricting to the predictions of a set of pre-trained ImageNet models, and then (2) repeating the `CONTAINS` task (cf. Section 2) on human annotators using the labels from (1) to identify the most reasonable ones.

#### B.1.1. PRE-FILTERING USING MODEL PREDICTIONS

We use the top-5 predictions of models with varying ImageNet (validation) accuracies (10 in total): `alexnet`, `resnet101`, `densenet161`, `resnet50`, `googlenet`, `efficientnet_b7` `inception_v3`, `vgg16`, `mobilenet_v2`, `wide_resnet50_2` (cf. Table 1) to identify a set of *potential labels*. Since model predictions tend to overlap, we end up with  $\sim 14$  potential labels per image on average (see full histogram in Figure 11). We always include the ImageNet label in the set of potential labels, even if it is absent in all the model predictions.

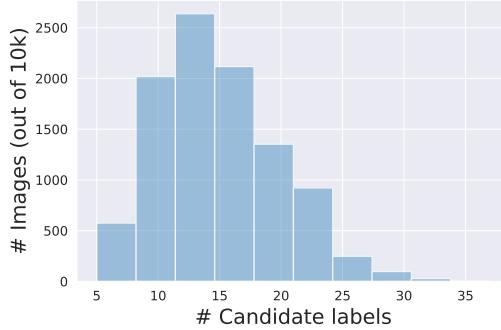


Figure 11: Distribution of labels per image obtained from the predictions of ImageNet-trained models (plus the ImageNet label). We present these labels (in separate grids) to annotators via the `CONTAINS` task (cf. Section 3.1) to identify a small set of relevant candidate labels for the classification task in Section 3.2.

#### B.1.2. MULTI-LABEL VALIDATION TASK

We then use human annotators to go through these potential labels and identify the most reasonable ones via the `CONTAINS` task. Recall that in `CONTAINS` task, annotators are shown a grid of images and asked to select the ones that contain an object corresponding to the specified query label (cf. Section 2). In our case, each image appears in multiple such grids—one for each potential label. By presenting these grids to multiple annotators, we can then obtain a selection frequency for every image-potential label pair, i.e., the number of annotators that perceive the label as being contained in the image (cf. Figure 7a). Using these selection frequencies, we identify the most relevant *candidate labels* for each image.

**Grid setup.** The grids used in our study contains 48 images, at least 5 of which are *controls*—obtained by randomly sampling from validation set images labeled as the query class. Along with the images, annotators are provided with a description of the query label in terms of (a) WordNet synsets and (b) the relevant Wikipedia link—see Figure 12 for an example. (Our MTurk interface is based on a modified version of the code made publicly available by Recht et al. (2019)<sup>5</sup>.) We find that a total of 3,934 grids suffice to obtain selection frequencies for all 10k images used in our analysis (w.r.t. all potential labels). Every grid was shown to 9 annotators, compensated \$0.20 per task.

**Quality control.** We filtered low-quality responses on a per-annotator and per-task basis. First, we completely omitted results from annotators who selected less than 20% of the control images on half or more of the tasks they completed: a

<sup>5</sup><https://github.com/modestyachts/ImageNetV2>

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](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.

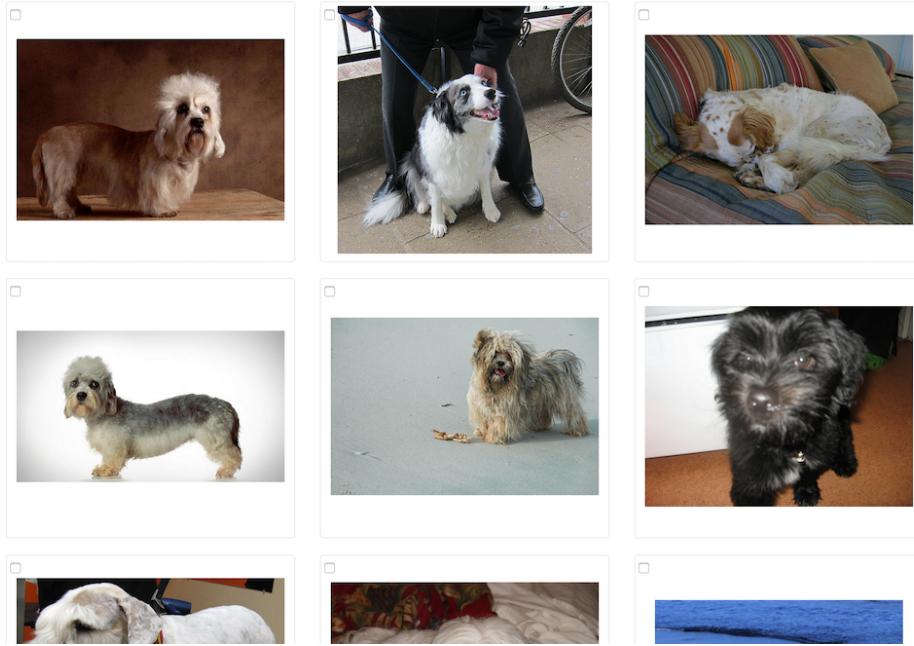


Figure 12: Sample interface of the CONTAINS task we use for label validation: annotators are shown a grid of 48 images and asked to select all images that correspond to a specific label (Section 3.1).

total of 10 annotators and the corresponding 513 tasks. Then we omitted tasks for which less than 40% of the controls were selected: at total of 3,104 tasks. Overall, we omitted 3,617 tasks in total out of the total 35,406. As a result, the selection frequency of some image-label pairs will be computed with fewer than 9 annotators.

#### B.1.3. FINAL CANDIDATE LABEL SELECTION

We then obtain the most relevant candidate labels by selecting the potential labels with high human selection frequency. To construct this set, we consider (in order):

1. The existing ImageNet label, irrespective of its selection frequency.
2. All the highly selected potential labels: for which annotator selection frequency is at least 0.5.
3. All potential labels with non-zero selection frequency that are semantically very different from the ImageNet label—so as to include labels that may correspond to different objects. Concretely, we select candidate labels that are more than 5 nodes away from the ImageNet label in the WordNet graph.
4. If an image has fewer than 5 candidates, we also consider other potential labels after sorting them based on their selection frequency (if non-zero).
5. To keep the number of candidates relatively small, we truncate the resulting set size to 6 if the excess labels have

selection frequencies lower than the ImageNet label, or the ImageNet label itself has selection frequency  $\leq 1/8$ . During this truncation, we explicitly ensure that the ImageNet label is retained.

In Figure 13, we visualize the distribution of number of candidate labels per image, over the set of images.

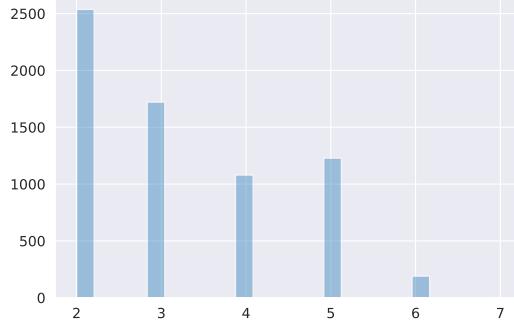


Figure 13: Distribution of the number of candidate labels used per image presented to annotators during the classification task in Section 3.2.

## B.2. Image classification

The candidate labels are then presented to annotators during the CLASSIFY task (cf. Section 3.2). Specifically, annotators are shown images, and their corresponding candidate labels and asked to select: a) all valid labels for that image, b) a label for the main object of the image—see Figure 14 for a sample task interface. We instruct annotators to pick multiple labels as valid, only if they correspond to different objects in the image and are not mutually exclusive. In particular, in case of confusion about a specific object label, we explicitly ask them to pick a single label making their best guess. Each task was presented to 9 annotators, compensated at \$0.08 per task.

**Images included.** We only conduct this experiment on images that annotators identified as having *at least* one candidate label outside the existing ImageNet label (based on experiment in Appendix B.1). To this end, we omitted images for which the ImageNet label was clearly the most likely: out of all the labels seen by 6 or more of the 9 annotators, the original label had more than double the selection frequency of any other class. Note that since we discard some tasks as part of quality control, it is possible that for some image-label pairs, we have the results of fewer than 9 annotators. Furthermore, we also omitted images which were not selected by any annotator as containing their ImageNet label (150 images total)—cf. Appendix Figure 26 for examples. These likely corresponds to labeling mistakes in the dataset creation process and do not reflect the systemic error we aim to study. The remaining 6,761 images that are part of our follow-up study have at least 1 label, in addition to the ImageNet label, that annotators think could be valid.

**Quality control.** Performing stringent quality checks for this task is challenging since we do not have ground truth annotations to compare against—which was after all the original task motivation. Thus, we instead perform basic sanity checks for quality control—we ignore tasks where annotators did not select *any* valid labels or selected a main label that they did not indicate as valid. In addition, if the tasks of specific annotators are consistently flagged based on these criteria (more than a third of the tasks), we ignore all their annotations. Overall, we omitted 1,269 out of the total 59,580 tasks.

The responses of multiple annotators are aggregated as described in Section 3.2.

**Select which labels appear in the image**  
 (Please read the instructions carefully as they are somewhat unusual)

**Task:**

- 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).

- 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
 <input checked="" type="radio"/> <input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="checkbox"/> Car <input checked="" type="checkbox"/> Car wheel <input type="checkbox"/> Truck
 <input type="radio"/>	<input type="radio"/>	<input checked="" type="checkbox"/> Fur <input type="checkbox"/> Wool <input checked="" type="checkbox"/> Fur coat
 <input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="checkbox"/> Collie <input type="checkbox"/> 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>

Figure 14: Screenshot of a sample image annotation task. (Section 3.2). Annotators are presented with an image and multiple candidate labels. They are asked to select all valid labels (selecting only one of mutually exclusive labels in the case of confusion) and indicate the main object of the image.

## C. Additional experimental results

### C.1. Multi-object Images

In Figure 15 we plot a histogram for the number of objects per image as indicated by our annotators, along with the most frequently co-occurring class pairs. For each object count, we plot the histogram of annotator selections in Figure 16—annotators seem to largely agree on the number of object present. Random multi-object images are shown in Figure 17.

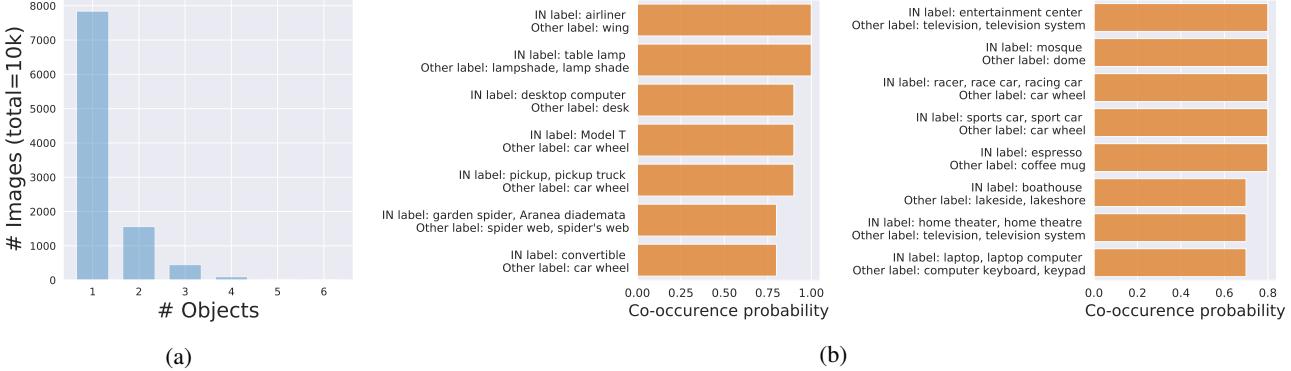


Figure 15: (a) Number of objects per image—more than a fifth of the images contains two or more objects from ImageNet classes (cf. Appendix Figure 17 examples). (b) Pairs of classes which consistently co-occur as distinct objects. Here, we visualize the top 15 ImageNet classes based on how often their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

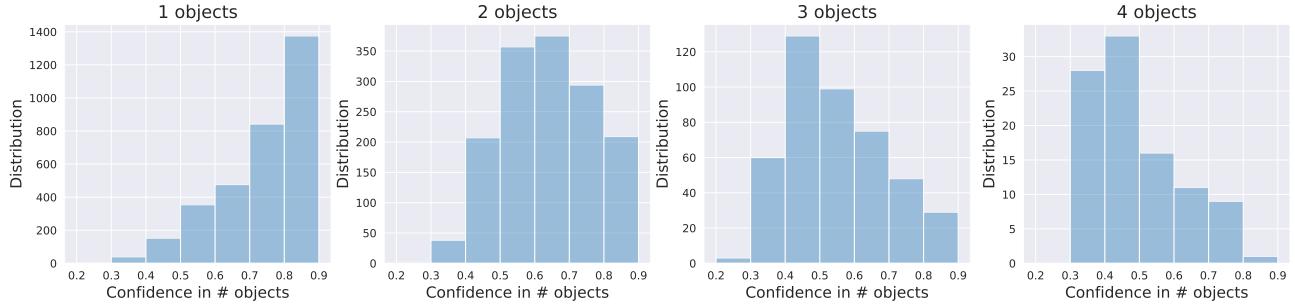


Figure 16: Annotator agreement for multi-object images. Recall that we determine the number of objects in an image based on a majority vote over annotators. Here, we define “confidence” as the fraction of annotators that make up that majority, relative to the total number of annotators shown the image (cf. Section 3.2). We visualize the distribution of annotator confidence, as a function of the number of image objects.

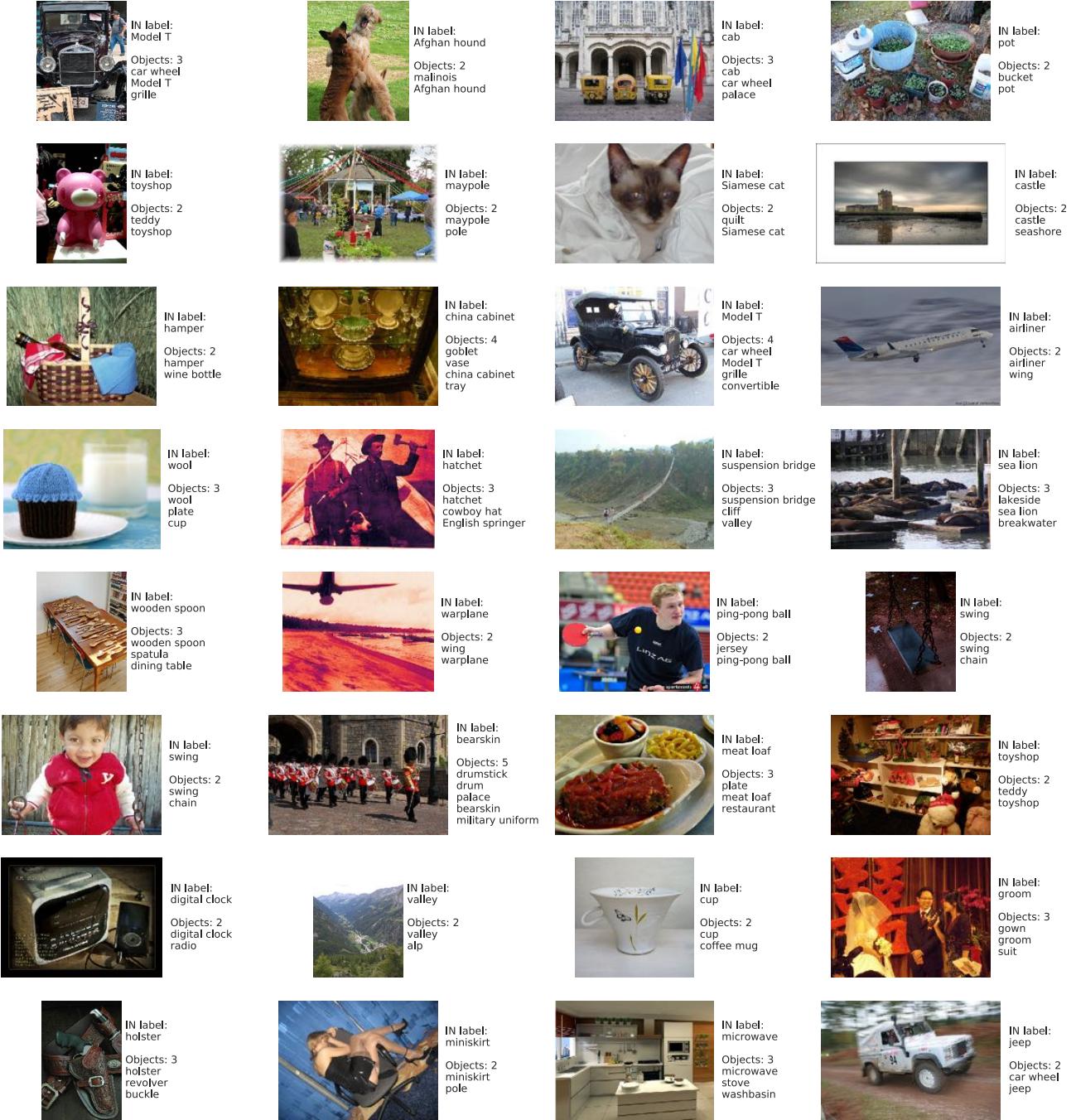


Figure 17: Sample ImageNet images with more than one valid label as per human annotators (cf. Section 3.1).

## From ImageNet to Image Classification: Contextualizing Progress on Benchmarks

In Figure 18, we visualize instances where the ImageNet label does not match with what annotators deem to be the “main object”. In many of these cases, we find that models perform at predicting the ImageNet label, even though the images contain other, more salient objects according to annotators—Figure 19.

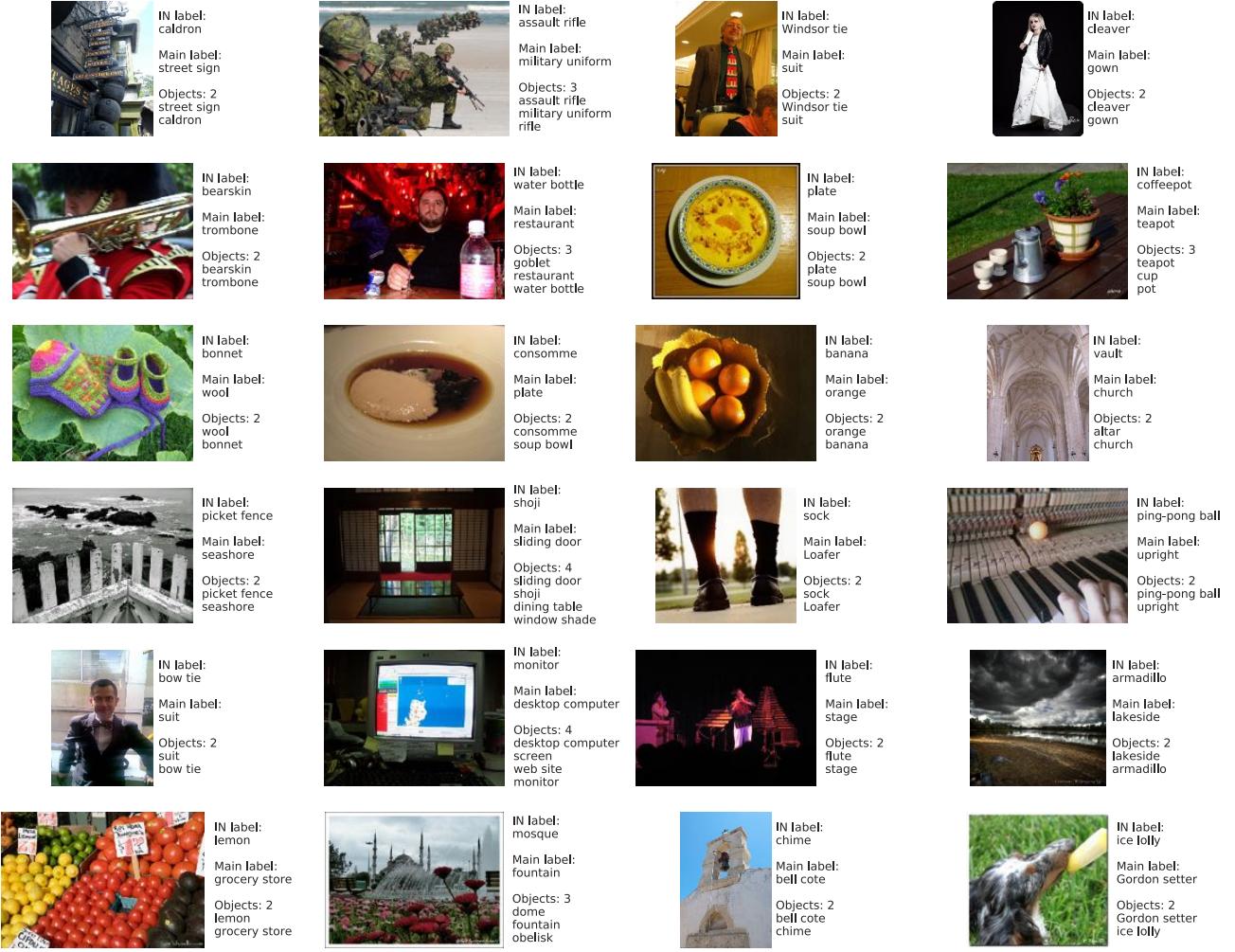


Figure 18: Sample images for which the main label as per annotators differs from the ImageNet label.

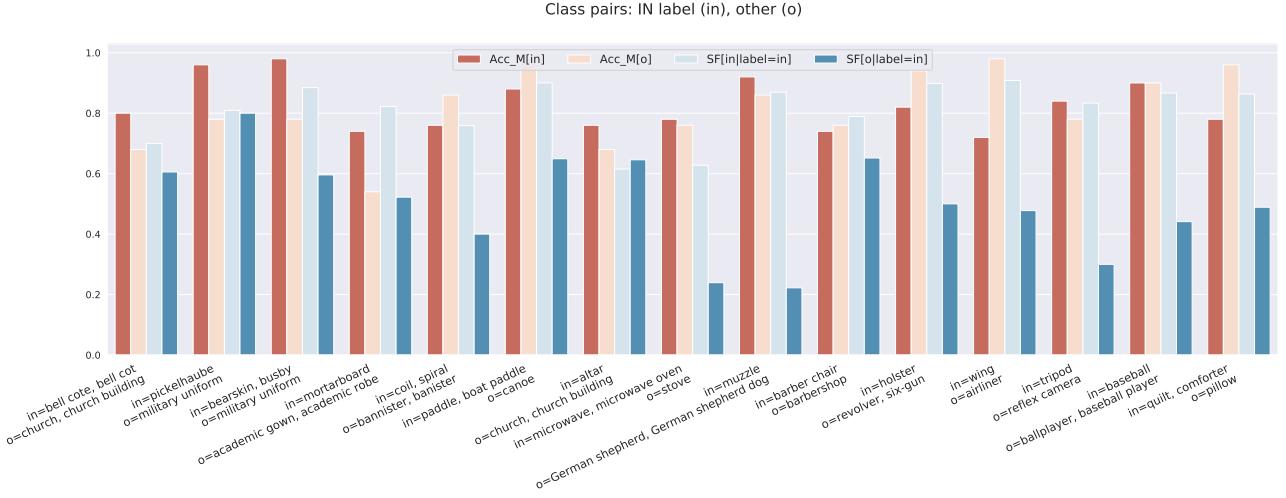


Figure 19: Classes for which human main label frequently differs from the ImageNet label. Here, although annotator selection frequency for the ImageNet label is high, the selection frequency for *another* class—which humans consider to be the main label—is also consistently high. Models still predict the ImageNet label, possibly by picking up on distinctive features of the objects.

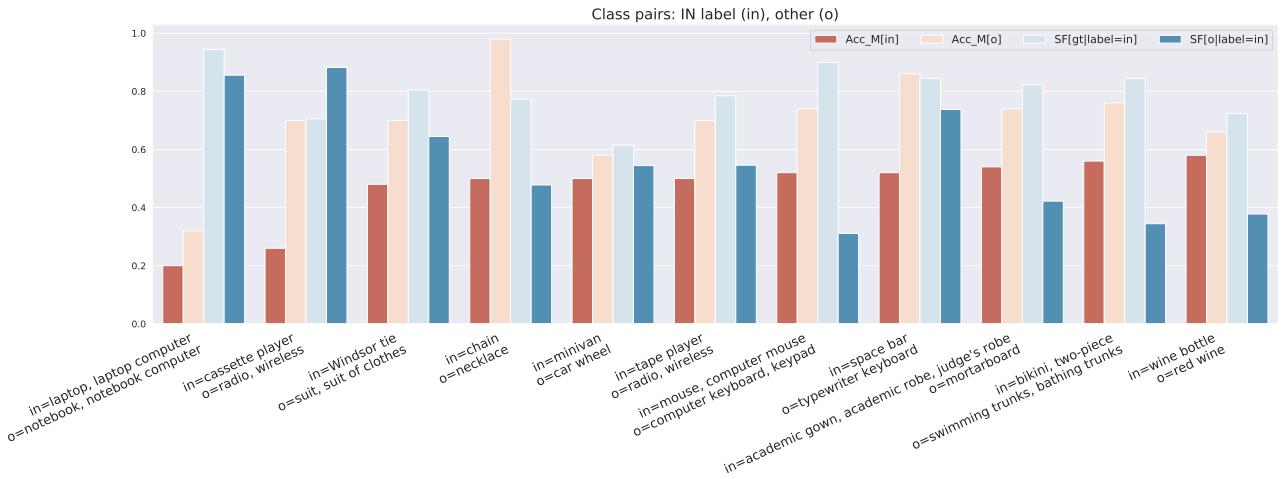


Figure 20: Classes where model accuracy is consistently low due to frequent object co-occurrences: in these cases, an object from the ImageNet class frequently co-occurs with (or is a sub-part of) objects from another class. Here, models seem to be unable to disambiguate the two classes completely, and thus perform poorly on one/both classes. Note that human selection frequency for the ImageNet class is high, indicating that an object from that class is present in the image.

**Top-5 accuracy in the multi-label context.** The issue of label ambiguity that can arise in multi-object images was noted by the creators of the ILSVRC challenge (Russakovsky et al., 2015). To tackle this issue, they proposed evaluating models based on top-5 accuracy. Essentially, a model is deemed correct if any of the top 5 predicted labels match the ImageNet label. We find that model top-5 accuracy is much higher than top-1 (or even our notion of multi-label) accuracy on multi-object images—see Figure 21. However, a priori, it is not obvious whether this increase is actually because of adjusting for model confusion between distinct objects. In fact, one would expect that properly accounting for such multi-object confusions should yield numbers similar to (and not markedly higher than) top-1 accuracy on single-object images.

To get a better understanding of this, we visualize the fraction of *top-5 corrections*—images for which the ImageNet label was not the top prediction of the model, but was in the top 5—that correspond to *different objects* in the image. Specifically, we only consider images where the top model prediction and ImageNet label were selected by annotators as: (a) present in the image and (b) corresponding to different objects. We observe that the fraction of top-5 corrections that correspond to multi-object images is relatively small—about 20% for more recent models. This suggests that top-5 accuracy may be overestimating model performance and, in a sense, masking model errors on single objects. Overall, these findings highlight the need for designing better performance metrics that reflect the underlying dataset structure.

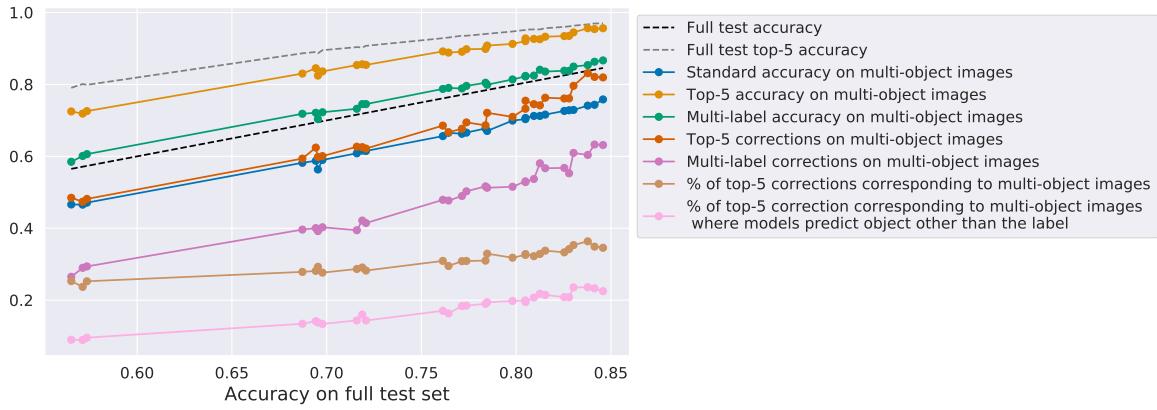


Figure 21: A closer look at top-5 accuracy: we visualize top-1, top-5 and multi-label (cf. Section 4.1) on multi-object images in ImageNet. We also measure the fraction of top-5 corrections (ImageNet label is not top model prediction, but is among top 5) that correspond to multi-object confusions—wherein the ImageNet label and top prediction belong to distinct image objects as per human annotators. We see that although top-5 accuracy is much higher than top-1, even for multi-object images, it may be overestimating model performance. In particular, a relatively small fraction of top-5 corrections actually correspond to the aforementioned multi-object images.

## C.2. Bias in label validation

In Figure 22 we plot the number of labels that were independently validated by at least two annotators each both for the CONTAINS task and the subsequent CLASSIFY task. In Figure 23 we plot model performance on classes that annotators often confuse with each other.

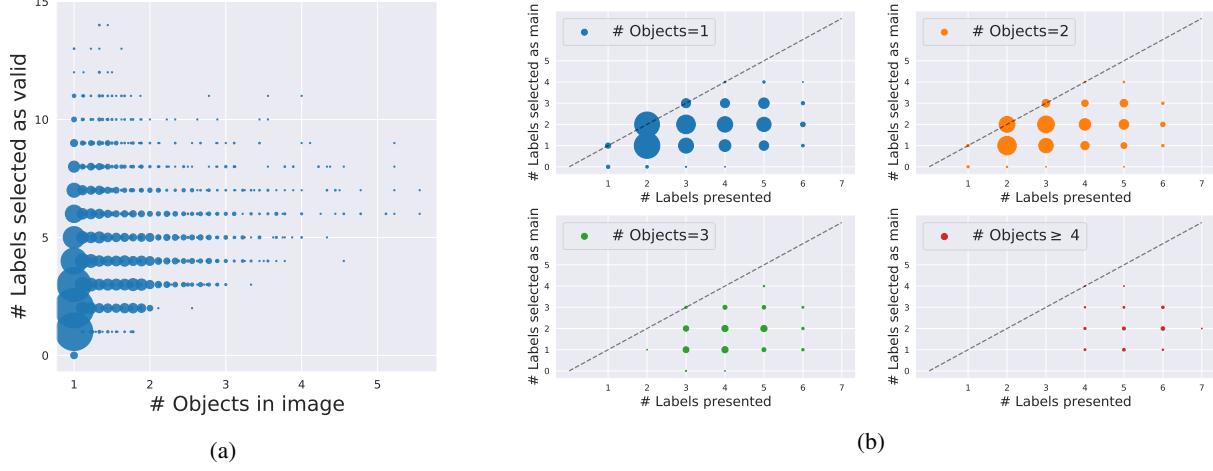


Figure 22: (a) Number of labels per image deemed valid by independent annotator groups in the CONTAINS task as a function of the average number of objects indicated to be present in the image during the CLASSIFY task; the dot size represents the number of images in each 2D bin. Even when annotators view an images as depicting a single object, they often collectively indicate 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 select significantly fewer labels when the task setup explicitly involves choosing between multiple labels *simultaneously*.

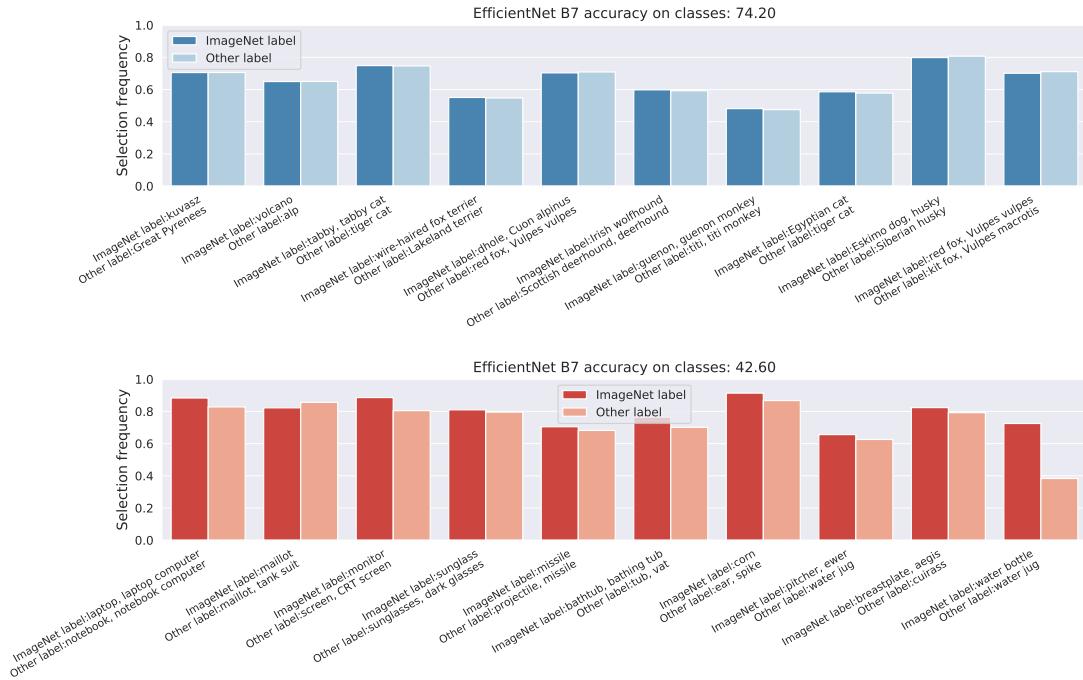


Figure 23: ImageNet class pairs for which annotators often deem both classes as valid in isolation. We visualize the top 10 pairs split based on the accuracy of EfficientNet B7 on these pairs being high (*top*) or low (*bottom*).

**Potential biases in selection frequency estimates.** In the course of obtaining fine-grained image annotations, we collect selection frequencies for several potential image labels (including the ImageNet label) using the `CONTAINS` task (cf. Section 3.1). Recall however, that during the ImageNet creation process, every image was already validated w.r.t. the ImageNet label (also via the `CONTAINS` task) by a different pool of annotators, and only images with high selection frequency actually made it into the dataset. This fact will result in a *bias* for our new selection frequency measurements (Engstrom et al., 2020). At a high level, if we measure a low selection frequency for the ImageNet label of an image, it is more likely that we are observing an underestimate, rather than the actual selection frequency being low. In order to understand whether this bias significantly affects our findings, we reproduce the relevant plots in Figure 24 using only a subset of workers (this should exacerbate the bias allowing us to detect it). We find however, that the difference is quite small, not changing any of the conclusions. Moreover, since most of our analysis is based on the per-image annotation task for which this specific bias does not apply, we can effectively ignore it in our study.

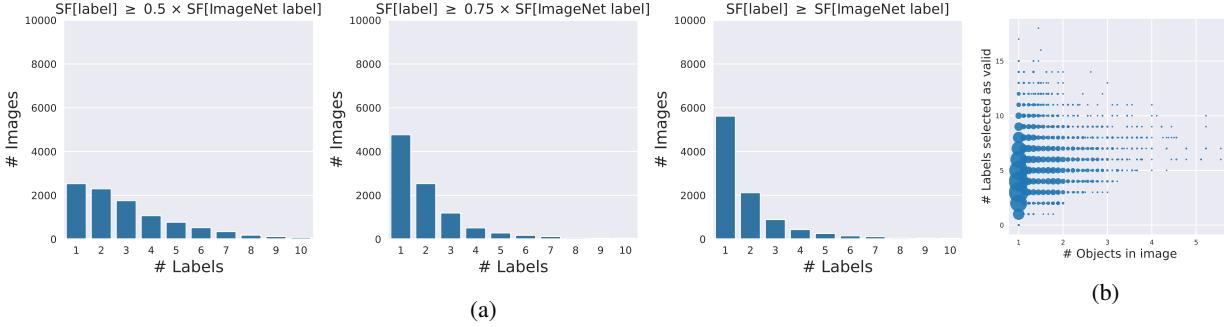


Figure 24: Effect of subsampling annotator population (5 instead of 9 annotators): (a) Number of labels annotators consider valid determined based on the selection frequency of a label relative to that of the ImageNet label. Even in this annotator subpopulation, for >70% of images, another label is still selected at least half as often as they select the ImageNet label (*leftmost*). (b) Number of labels that at least one (of five) annotators selected as valid for an image (cf. Section 3.1) versus the number of objects in the image (cf. Section 3.2). (Dot size is proportional to the number of images in each 2D bin.) Even when annotators consider the image as containing only a single object, they often select multiple labels as valid.

**Selection frequency and model accuracy.** In typical dataset creation pipelines, selection frequency (and other similar metrics) is often used to filter out images that are not easily recognizable by humans. We now consider how the selection frequency of a class relates to model accuracy on that class—see Figure 25. While we see that, in general, classes with high human selection frequency are also easier for models to classify, this is not uniformly true—in particular, there seem to be classes that significantly deviate from this trend. For instance, multi-object images would have high human selection frequency for the ImageNet label, but make predicting the ImageNet label challenging

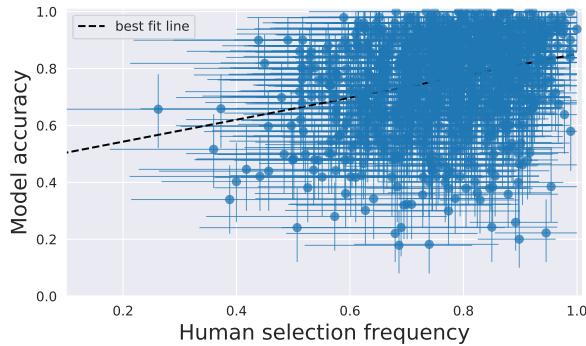


Figure 25: Relationship between per-class selection frequency and model (ResNet-50) accuracy. We observe that while in general higher selection frequency does correlate with better accuracy, this is not uniformly true. This suggests that using selection frequency as a proxy for how easy an image is in terms of object recognition may not be perfect—especially if the dataset contains fine-grained classes or multi-object images.

### C.3. Mislabeled examples

In the course of our human studies in Section 3.1, we also identify a set of possibly mislabeled ImageNet images. Specifically, we find images for which:

- Selection frequency for the ImageNet label is 0, i.e., no annotator selected the label to be contained in the image (cf. Section 3.1). We identify 150 (of 10k) such images— cf. Appendix Figure 26 for examples.
- The ImageNet label was not selected at all (for any object) during the detailed image annotation phase in Section 3.2. We identify 119 (of 10k) such images— cf. Appendix Figure 27 for examples.

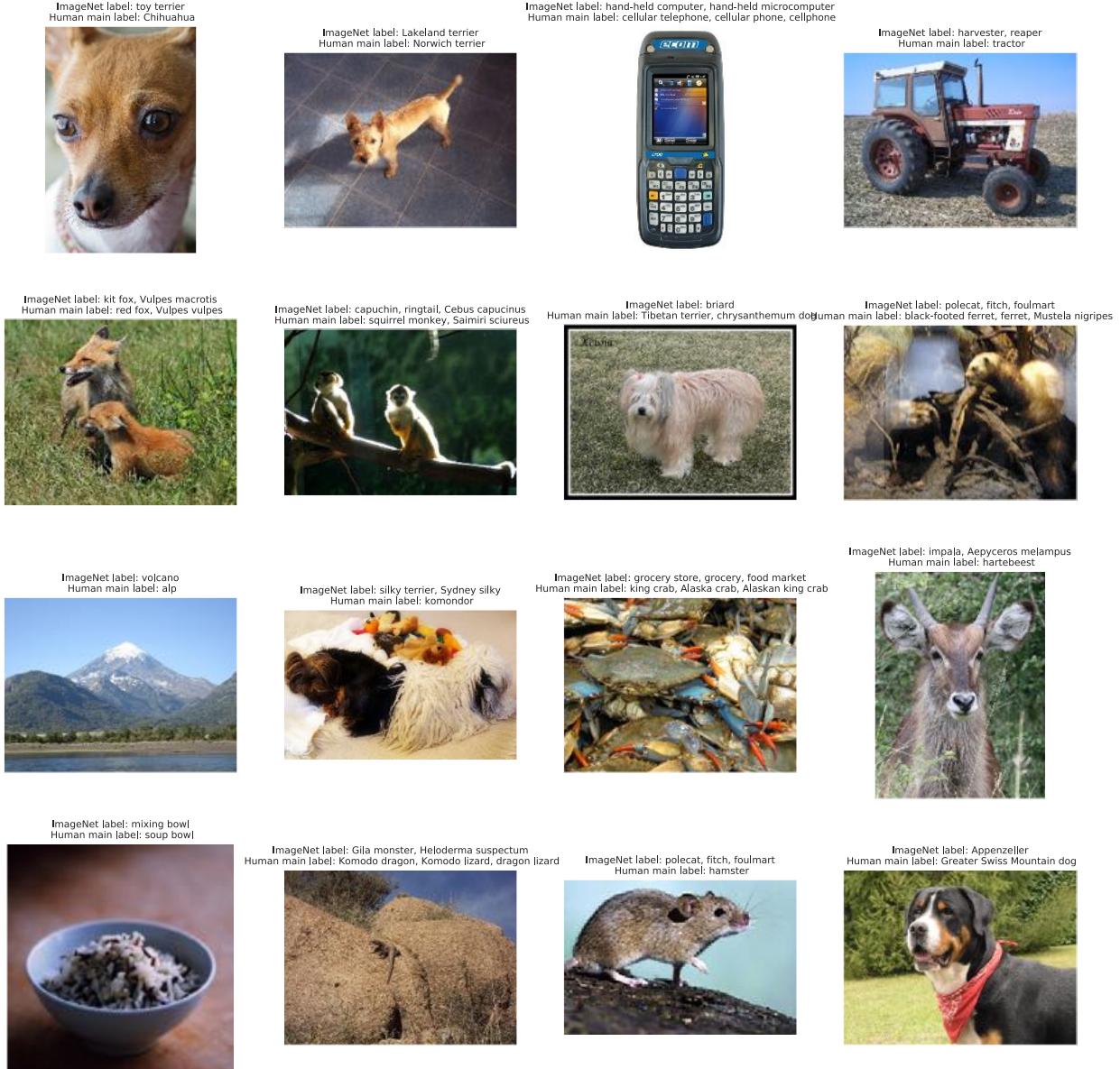


Figure 26: Possibly mislabeled images: human selection frequency for the ImageNet label is 0 (cf. Section 3.1). Also depicted is the label most frequently selected by the annotators as contained in the image (*sel*).

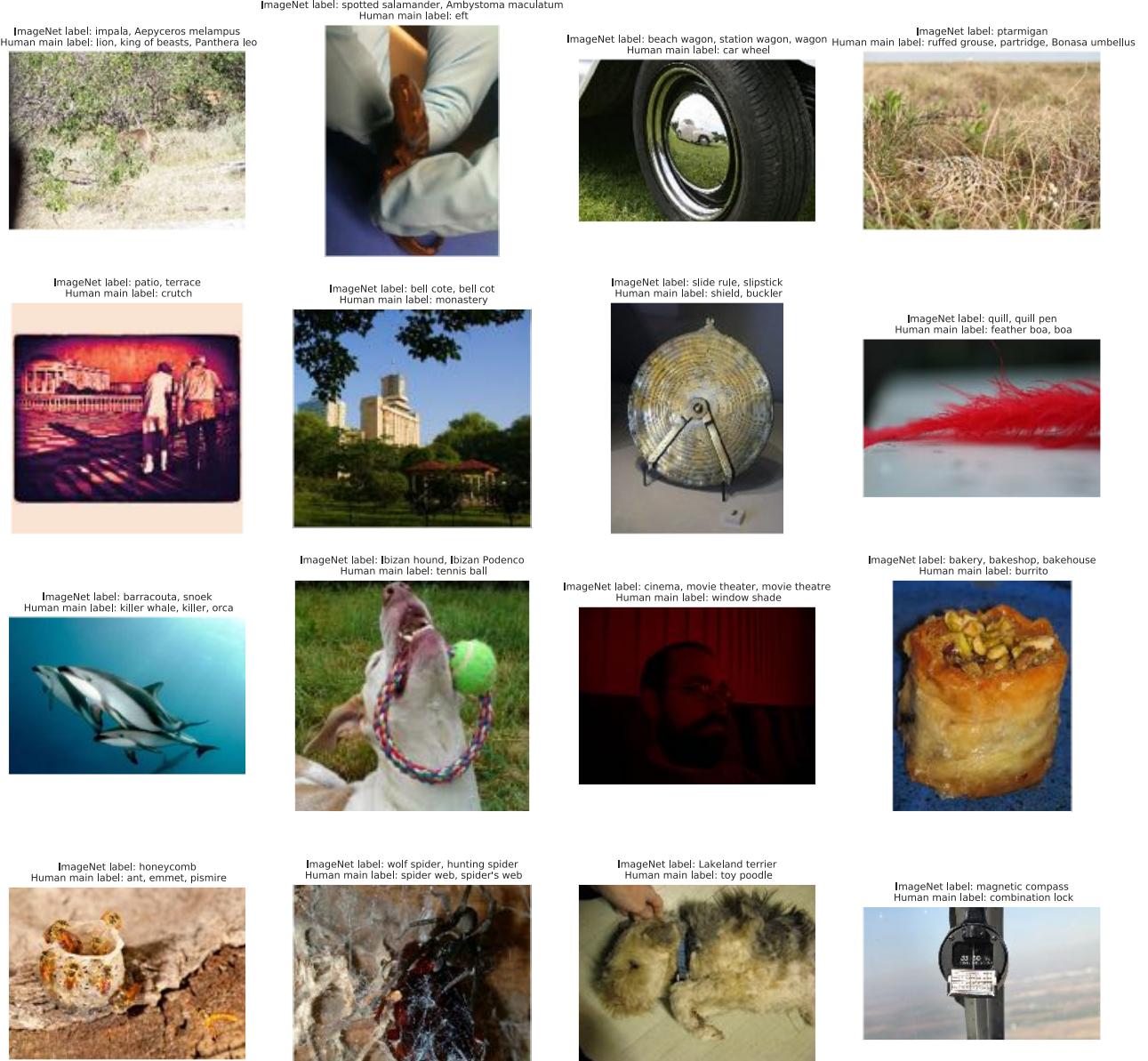


Figure 27: Possibly mislabeled images: ImageNet label is not selected by any of the annotators during fine-grained image annotation process described in Section 3.2. Also shown in the title is label that was most frequently selected by the annotators as denoting the main object in the image (*sel*).

#### C.4. Confusion Matrices

The  $(i, j)$ th entry of the human/model confusion matrix denotes how often an image with ImageNet label  $i$  is predicted as class  $j$ . We consider the model prediction to simply be the top-1 label. We determine the “human prediction” in two ways, as the class: (1) with highest annotator selection frequency in the CONTAINS task, and (2) which is the most likely choice for the main label in the CLASSIFY task. In Appendix Figure 28 we compare model and human confusion matrices (for both notions of human prediction). Unless otherwise specified, all other confusion matrices in the paper are based on the main label (using method (2)).

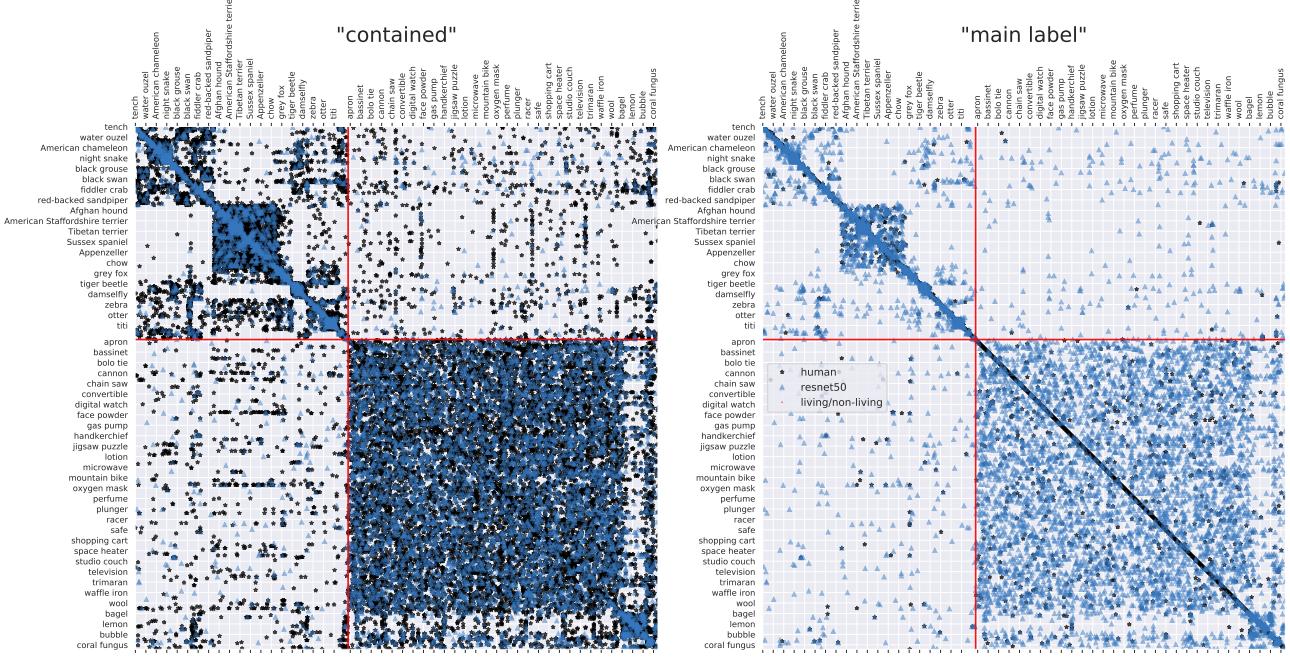


Figure 28: Comparison of model (blue; ResNet-50) and human (black) confusion matrices for all 1000 ImageNet classes. At a high-level, model and human confusion patterns seem somewhat aligned although humans seem to be particularly worse when it comes to fine-grained classes.

We compare the inter-superclass confusion matrices for humans and various models in Appendix Figure 29. We see that as models get better, their confusion patterns look similar to human annotators. Moreover, as we saw previously in Figure 28, there are blocks of superclasses where confusion seems particularly prominent—likely due to frequent object co-occurrences (cf. Figure 10b). We find that intra-superclass confusions are more prominent for humans—see Appendix Figure 30). This is in line with our findings from Section 3.1, where we observe that human annotators often select multiple labels for an image, even when they think it contains one object.

## From ImageNet to Image Classification: Contextualizing Progress on Benchmarks

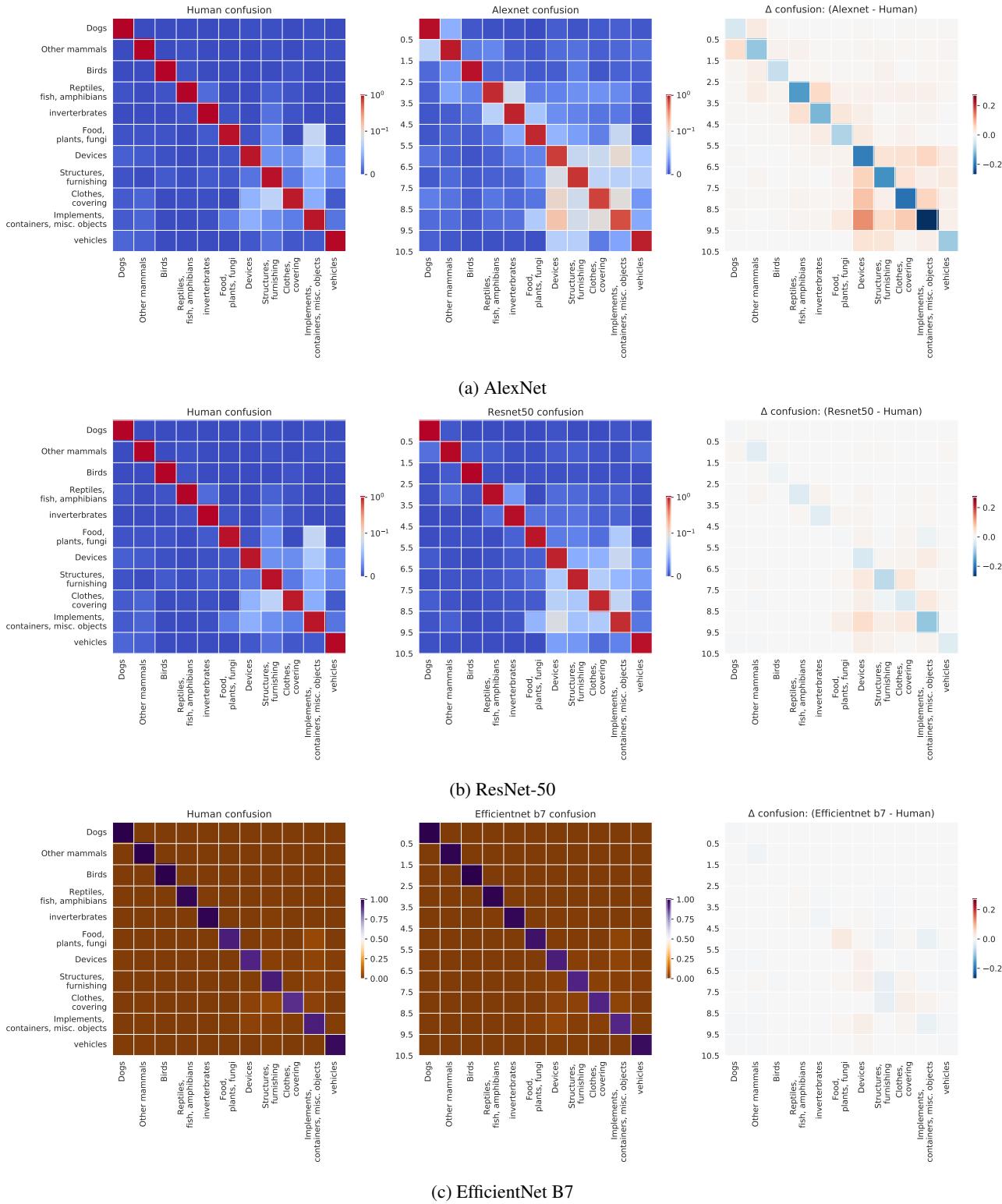
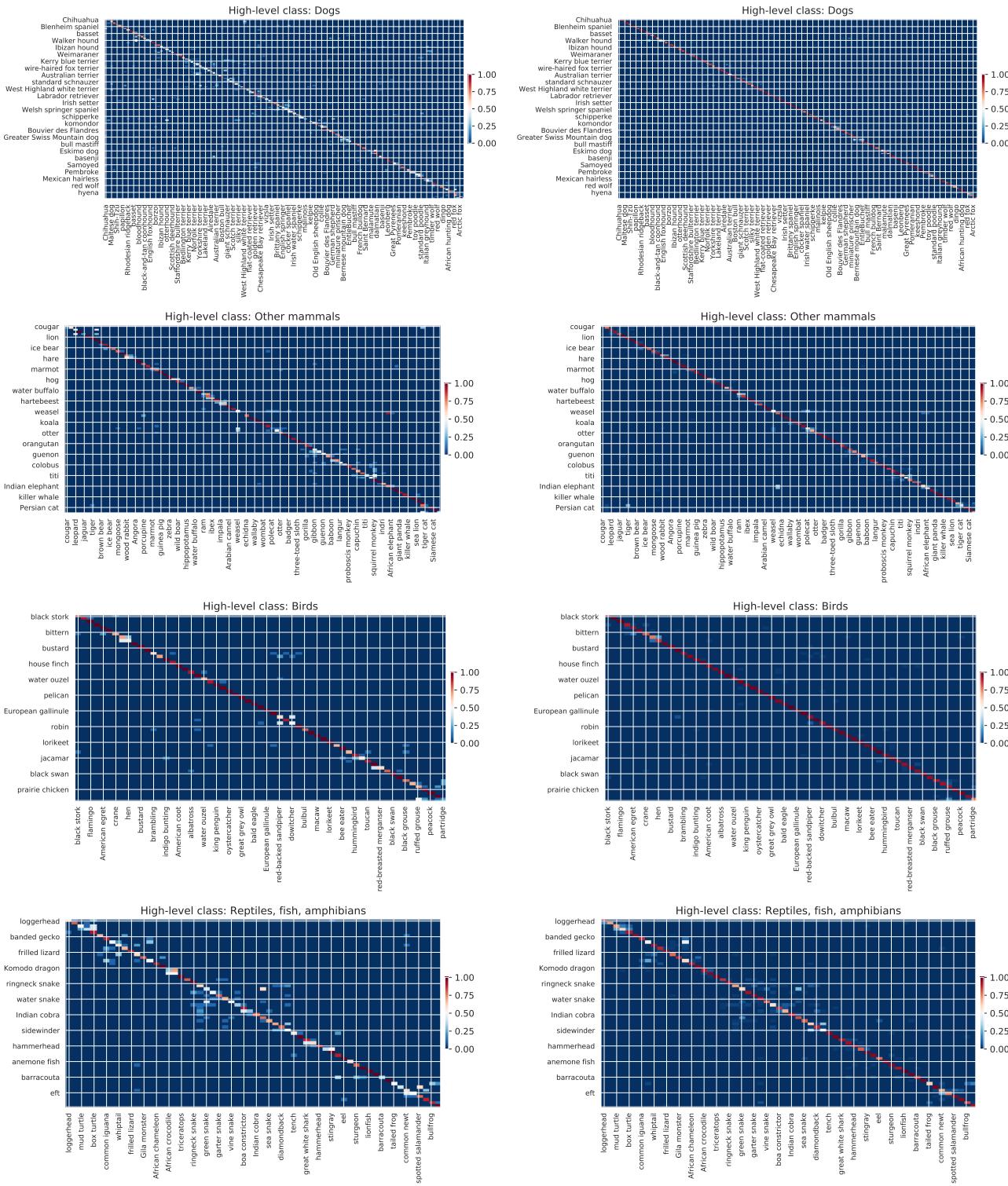
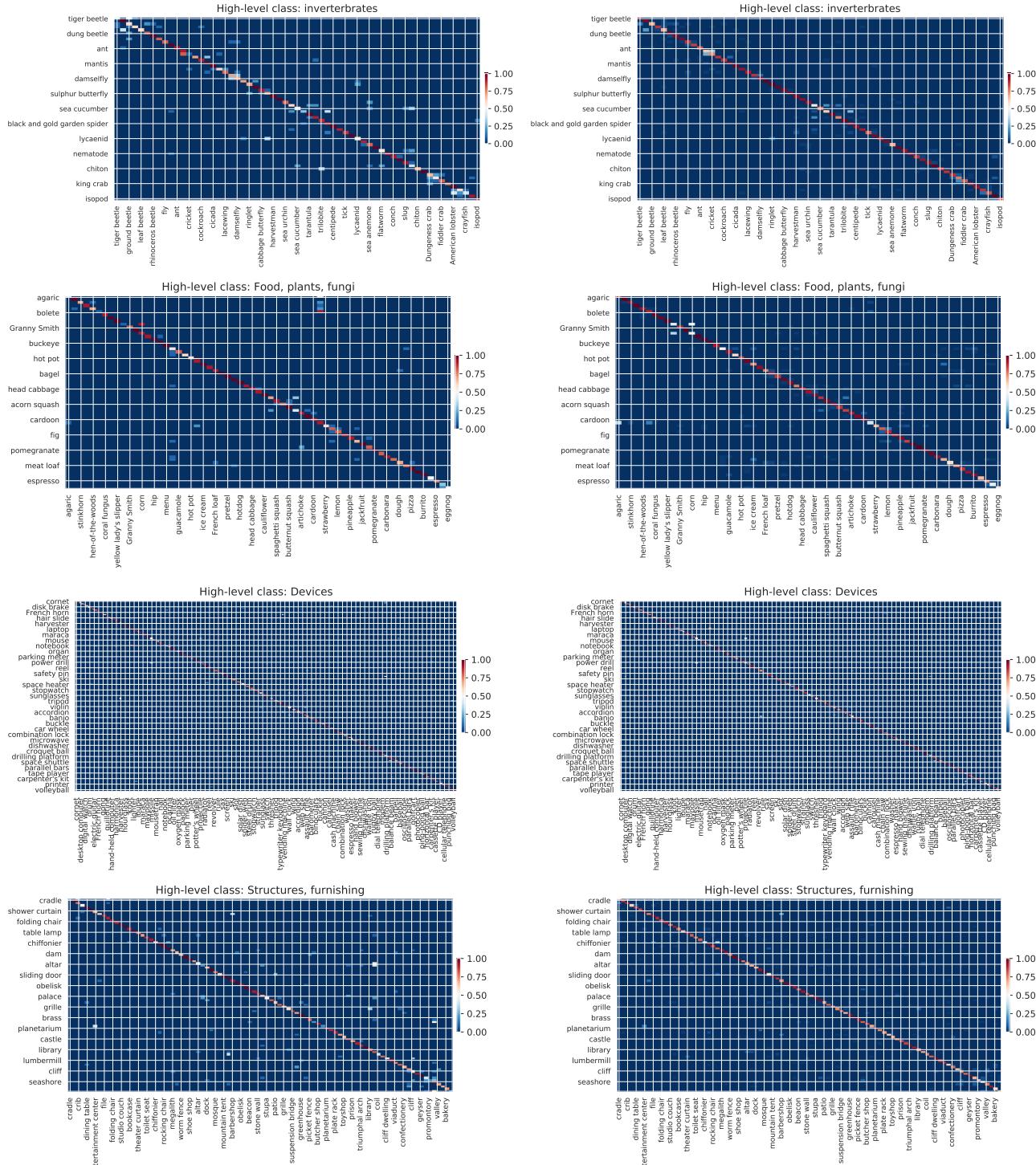


Figure 29: Inter-superclass confusion matrices for models and humans. We observe that as models get more accurate, their confusion also tends to align better with humans. In fact, more recent models seem to be better than humans, especially when it comes to fine-grained classes. Moreover, we see that for these models, the confusion patterns seem to mirror the object co-occurrences in Figure 10b. This suggests that part of errors of current models may stem from (legitimate) confusions due to multi-label images.

# From ImageNet to Image Classification: Contextualizing Progress on Benchmarks



## From ImageNet to Image Classification: Contextualizing Progress on Benchmarks



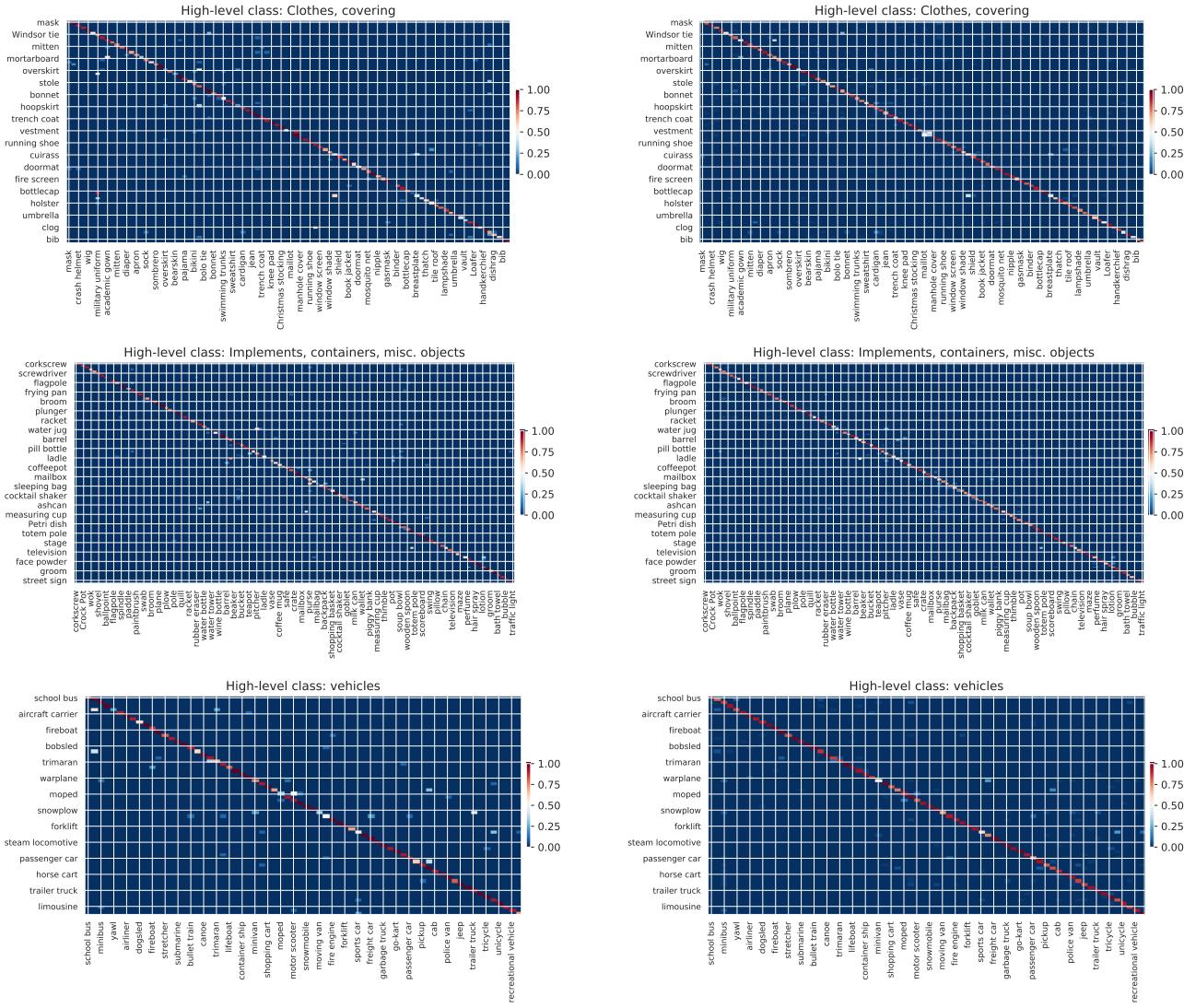


Figure 30: Intra-superclass confusion matrices for (left) humans and (right) an EfficientNet B7.