Contemplating Real-World Object Classification

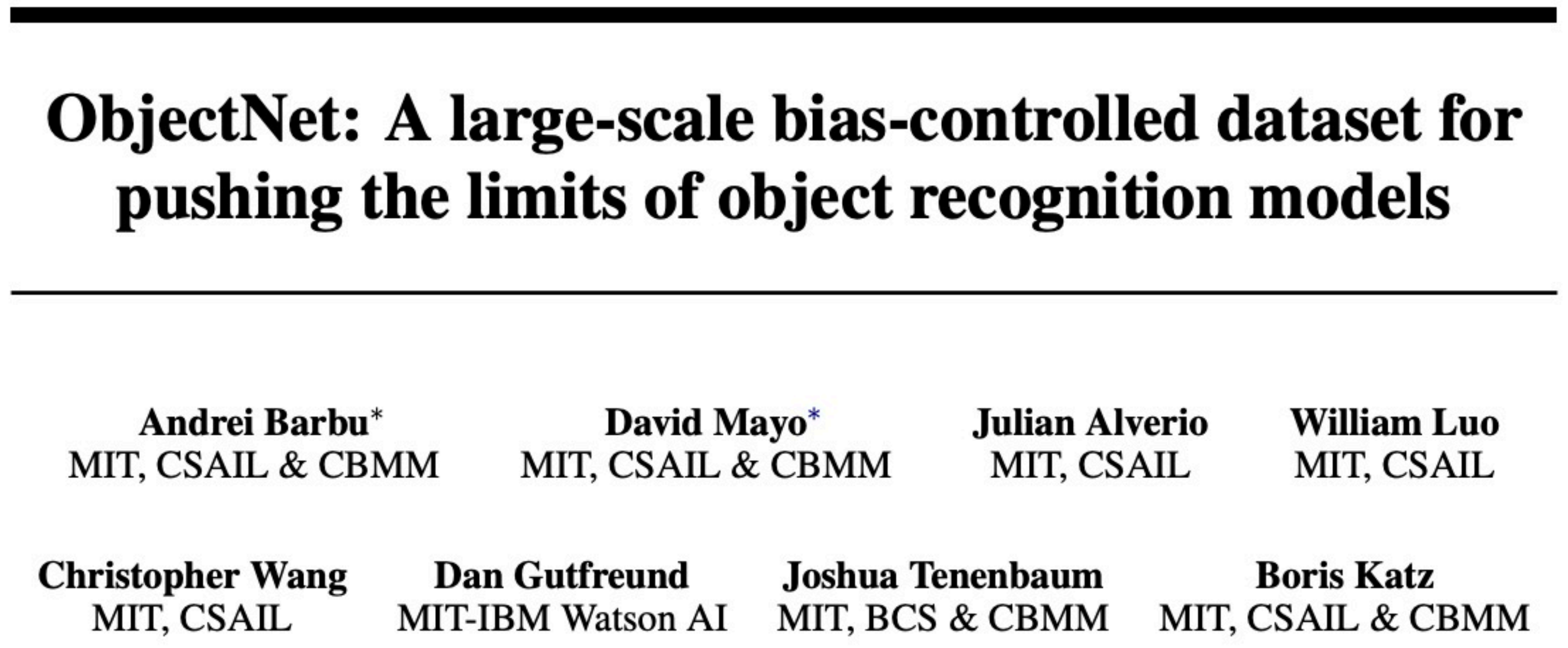
Ali Borji (aliborji@gmail.com)

ICLR

Motivation

- Deep object recognition models have been very successful over benchmark datasets such as ImageNet. **How accurate and robust are they to distribution shifts arising from natural and synthetic variations in datasets?**
- Here, we reanalyze the ObjectNet dataset recently proposed by Barbu et al. (NIPS 2019) containing objects in daily life situations.
- We find that applying deep models to the **isolated objects, rather than the entire scene as is done in the original paper, results in around 20-30% performance improvement.** Relative to the numbers reported in Barbu et al., around 10-15% of the performance loss is recovered, without any test time data augmentation.
- We also investigate the **robustness of models against synthetic image perturbations** such as geometric transformations (e.g., scale, rotation, translation), natural image distortions (e.g., impulse noise, blur) as well as adversarial attacks (e.g., FGSM and PGD-5).
- **Code and data are available at https://github.com/aliborji/ObjectNetReanalysis.git**

ObjectNet Dataset



ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models

Andrei Barbu*
MIT, CSAIL & CBMM

David Mayo*
MIT, CSAIL & CBMM

Julian Alverio
MIT, CSAIL

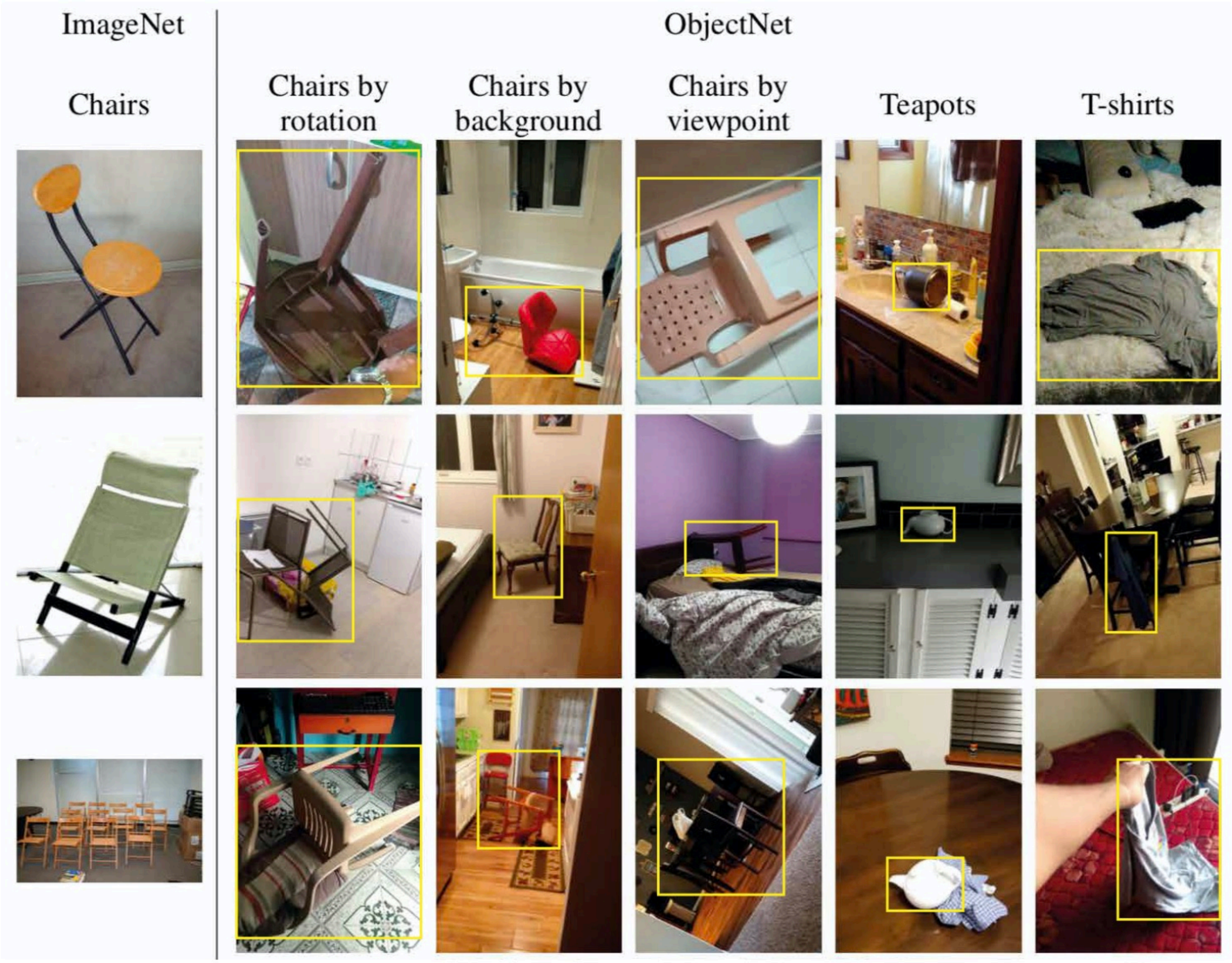
William Luo
MIT, CSAIL

Christopher Wang
MIT, CSAIL

Dan Gutfreund
MIT-IBM Watson AI

Joshua Tenenbaum
MIT, BCS & CBMM

Boris Katz
MIT, CSAIL & CBMM



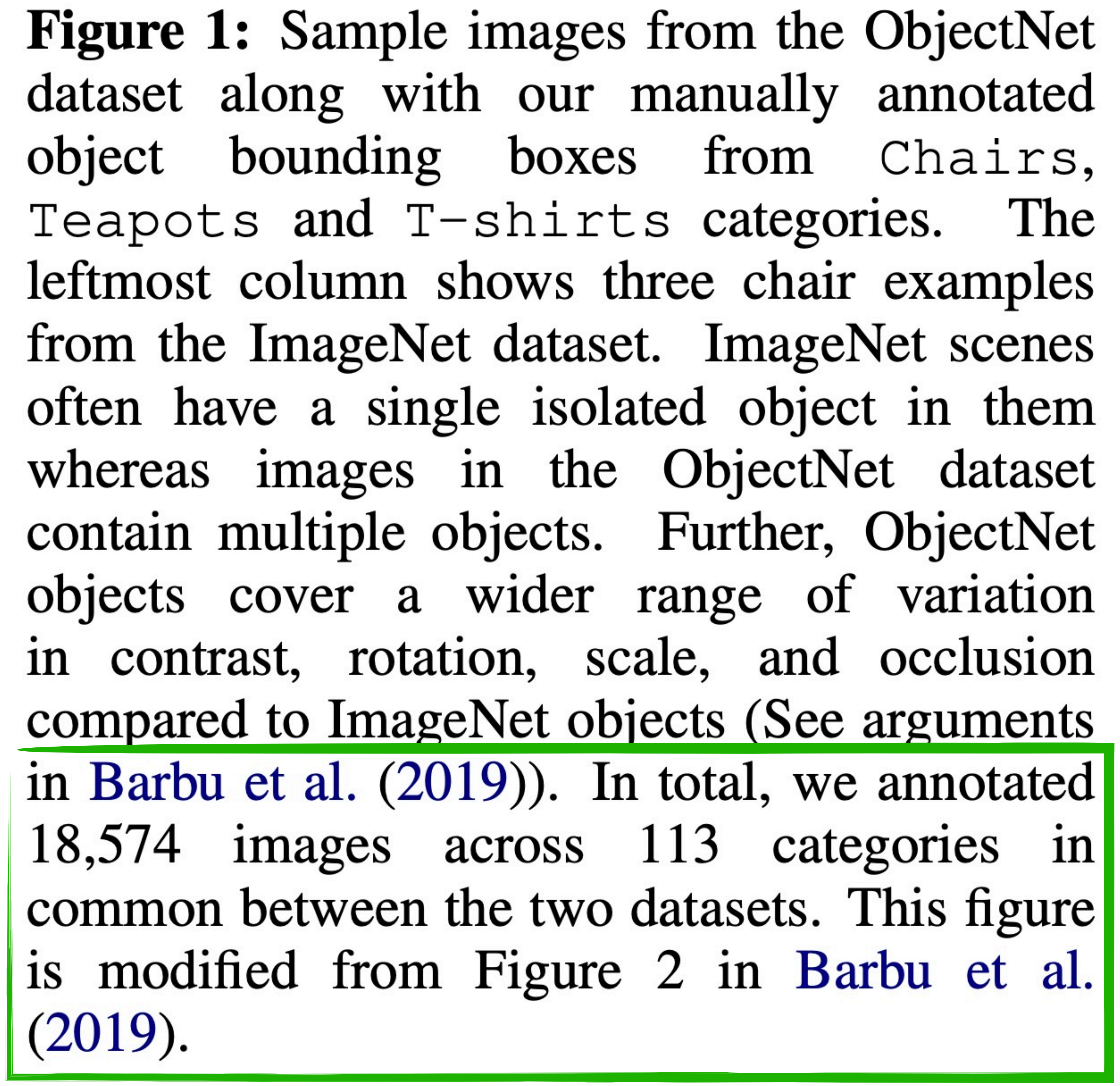
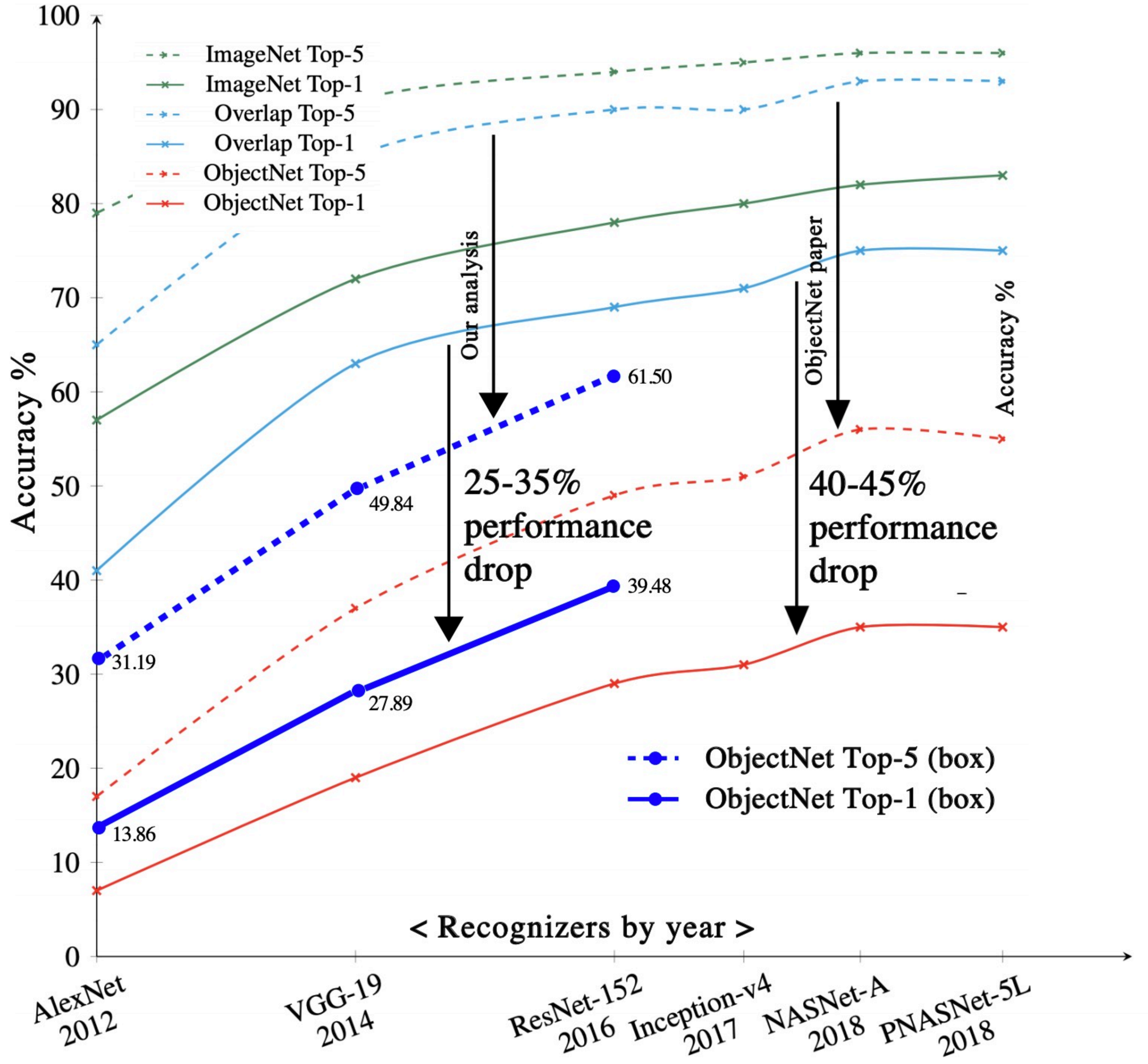


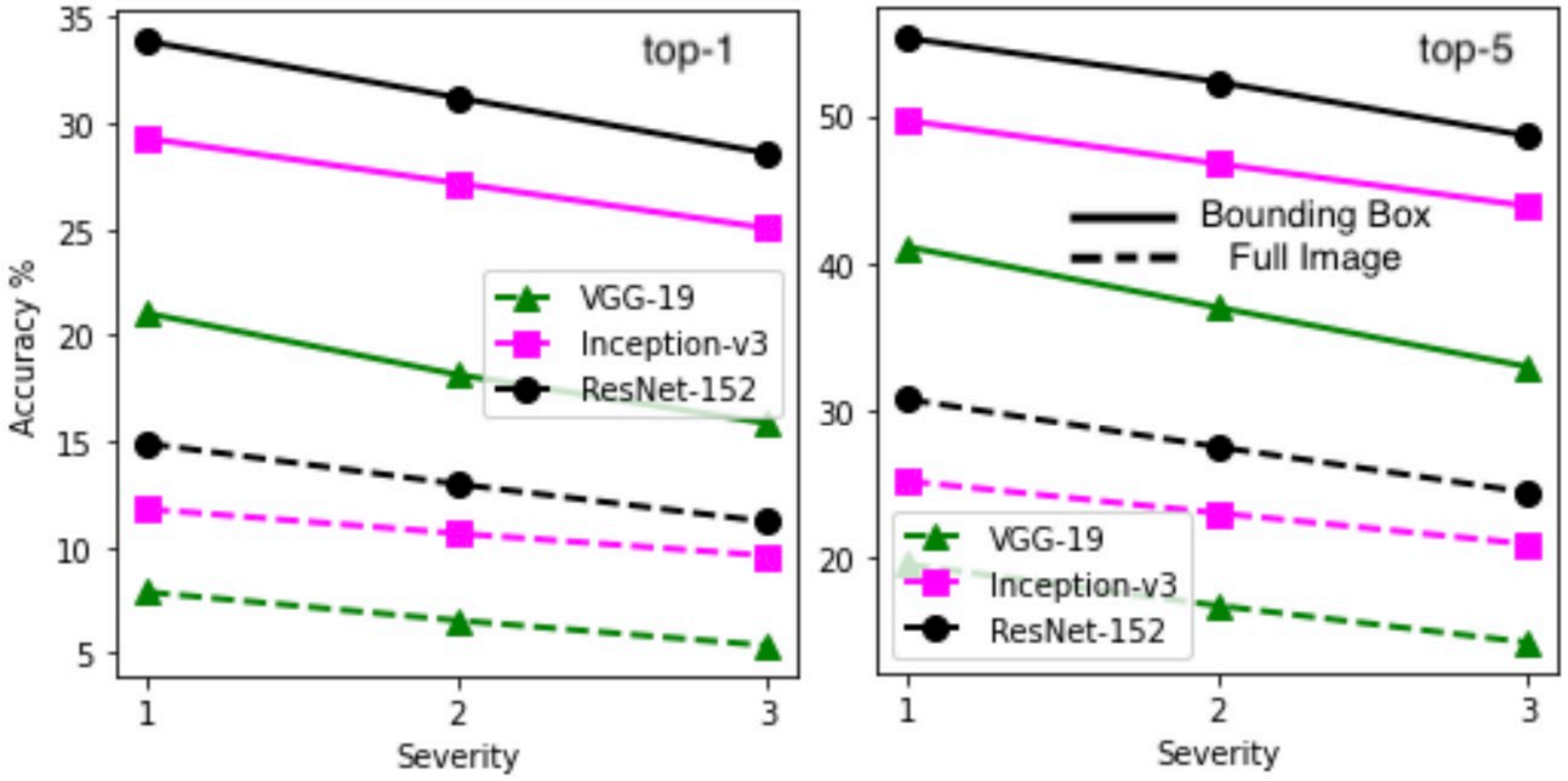
Figure 1: Sample images from the ObjectNet dataset along with our manually annotated object bounding boxes from `Chairs`, `Teapots` and `T-shirts` categories. The leftmost column shows three chair examples from the ImageNet dataset. ImageNet scenes often have a single isolated object in them whereas images in the ObjectNet dataset contain multiple objects. Further, ObjectNet objects cover a wider range of variation in contrast, rotation, scale, and occlusion compared to ImageNet objects (See arguments in Barbu et al. (2019)). In total, we annotated 18,574 images across 113 categories in common between the two datasets. This figure is modified from Figure 2 in Barbu et al. (2019).

Accuracy And Robustness Against Natural Distribution Shifts

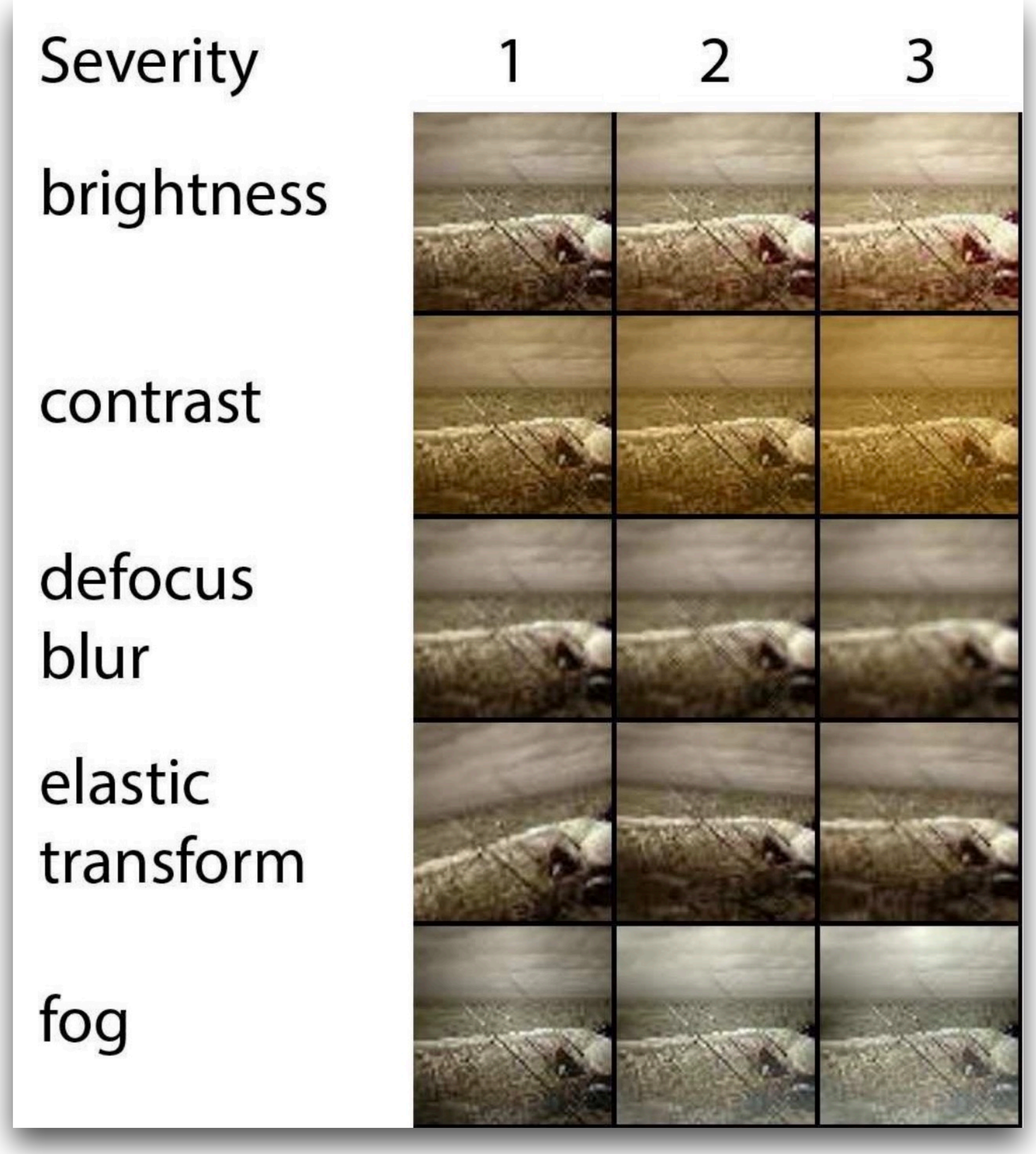
- Performance of deep object recognition models on ObjectNet dataset.
- Feeding object boxes to models instead of the entire scene improves the accuracy about 10-15%.
- The right panel shows the results using our code (thus leading to a fair comparison). Performance improvement is more significant now which is ~ 20-30%.



Robustness Against Common Image Corruptions

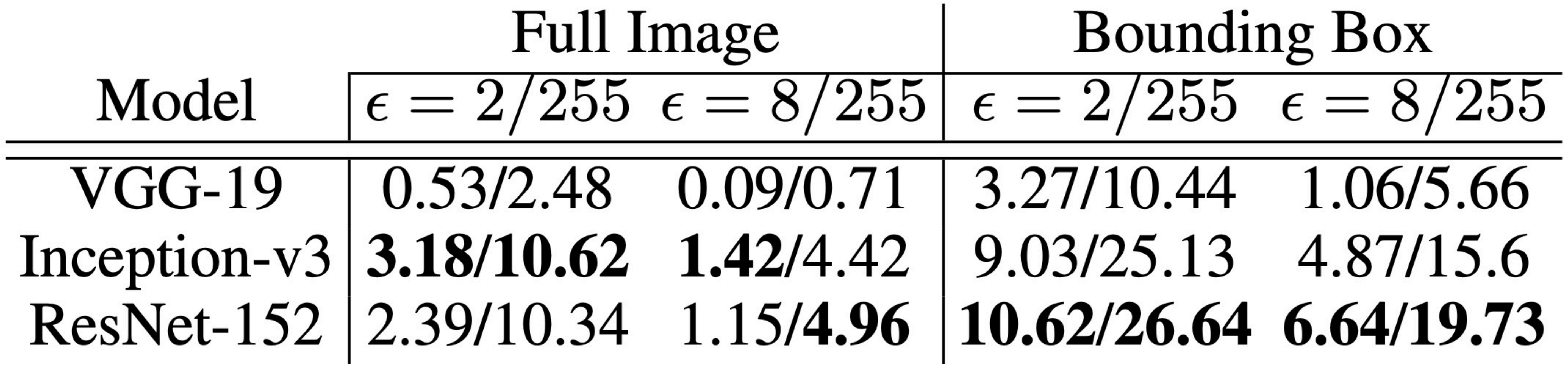


Average top-1 and top-5 accuracy of models over 1130 images from the ObjectNet dataset corrupted by 14 natural image distortions.



Robustness Against Adversarial Perturbations

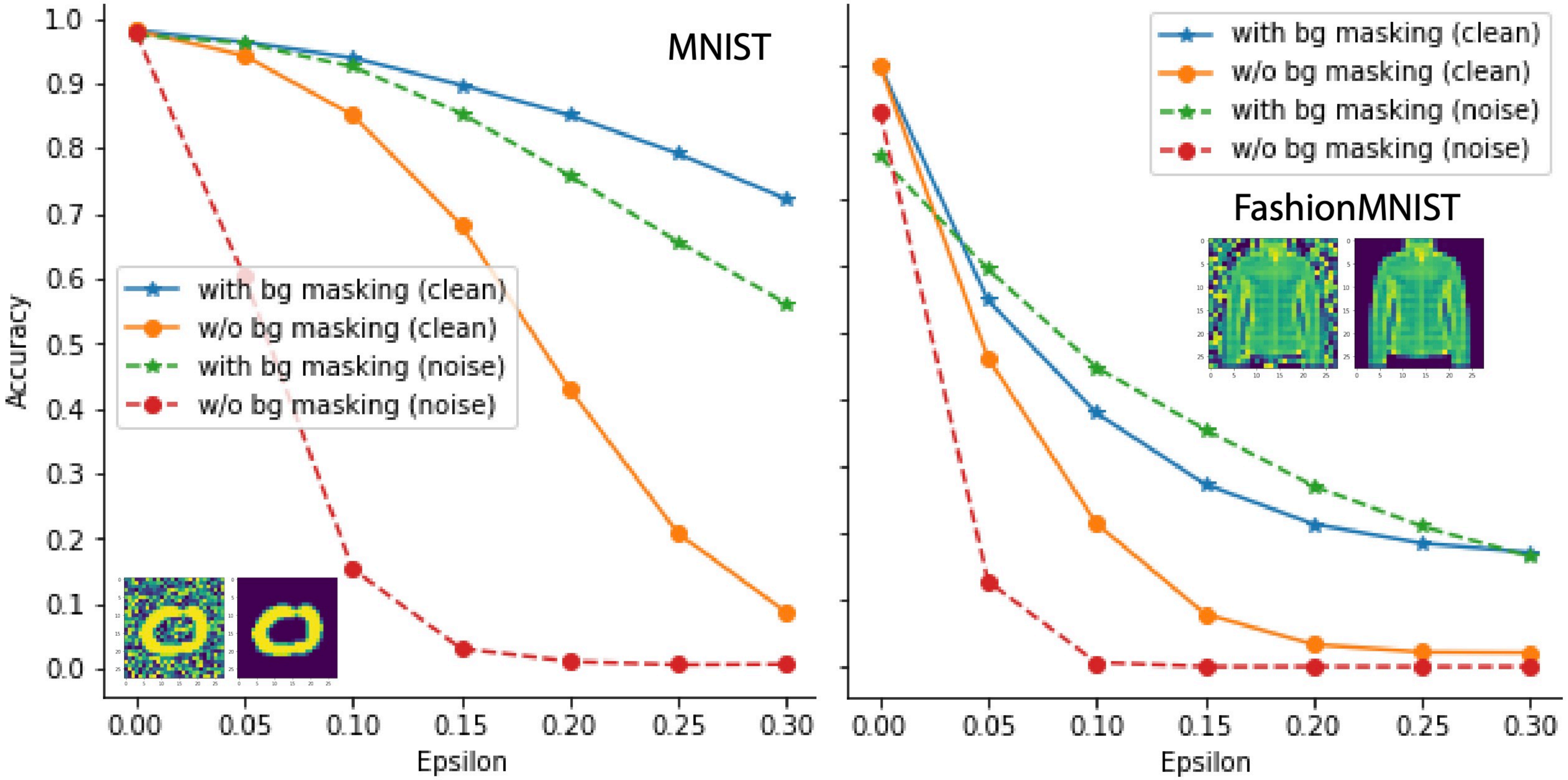
Robust accuracy (top-1/top-5) against FGSM attack.



	Full Image		Bounding Box	
Model	$\epsilon = 2/255$	$\epsilon = 8/255$	$\epsilon = 2/255$	$\epsilon = 8/255$
VGG-19	0.53/2.48	0.09/0.71	3.27/10.44	1.06/5.66
Inception-v3	**3.18/10.62**	**1.42**/4.42	9.03/25.13	4.87/15.6
ResNet-152	2.39/10.34	1.15/**4.96**	**10.62/26.64**	**6.64/19.73**

** We find that models are more resilient against the FGSM attack when applied to the bounding box than the full image.

The Influence Of The Surrounding Context On Robustness



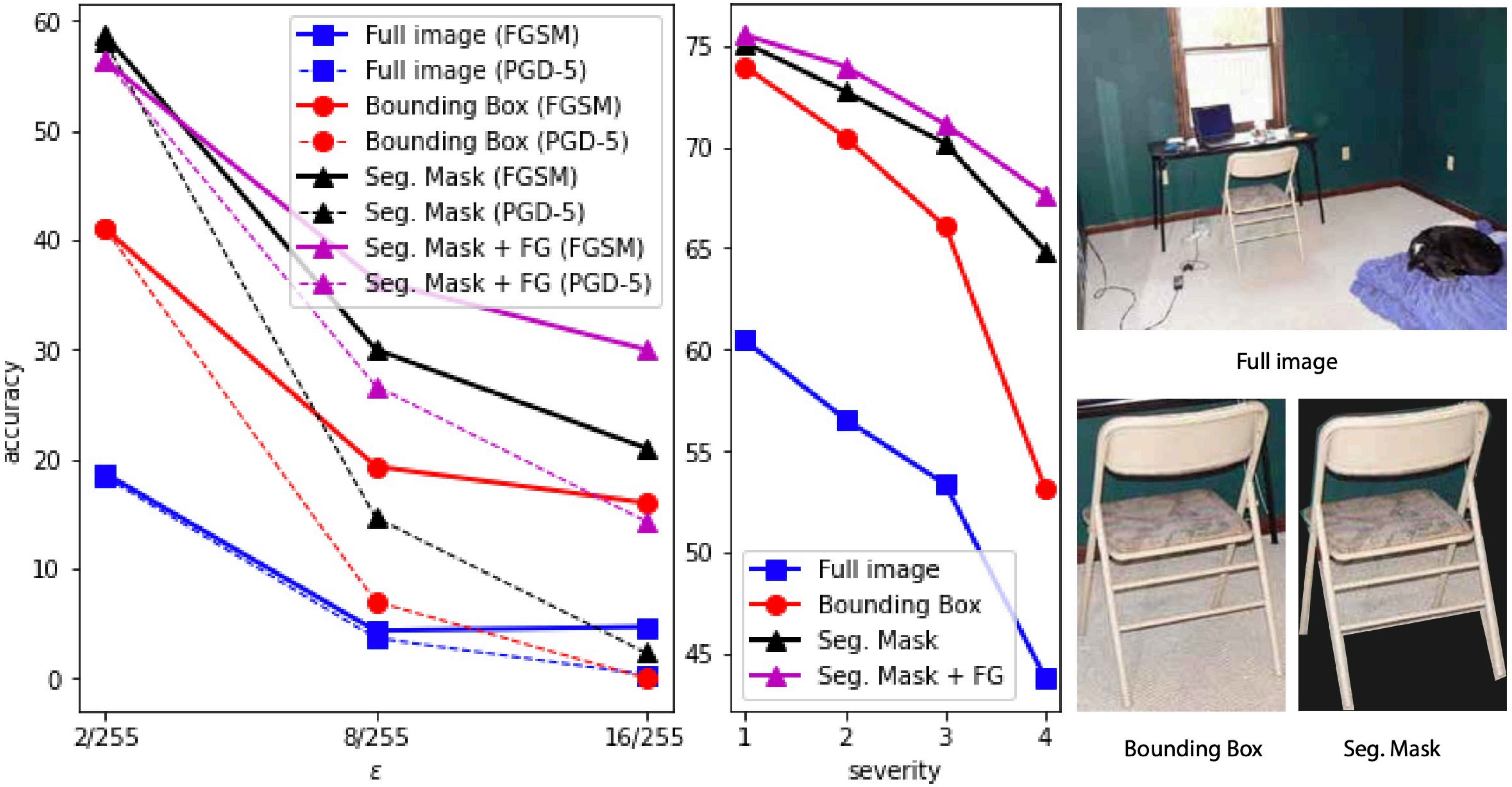
The effect of background subtraction (a.k.a foreground detection) on adversarial robustness (here against the FGSM attack). Two models are trained and tested on clean and noisy data from MNIST (left) and FashionMNIST (right) datasets. In the noise case, the object is overlaid in a white noise field (no noise on the object itself).

The Influence Of The Surrounding Context On Robustness (Cnt’d)

- Images from ten classes of the MS COCO dataset including chair, car, book, bottle, dinning table, umbrella, boat, motorcycle, sheep, and cow (one object per image; 100 images per category; 1000 images in total)

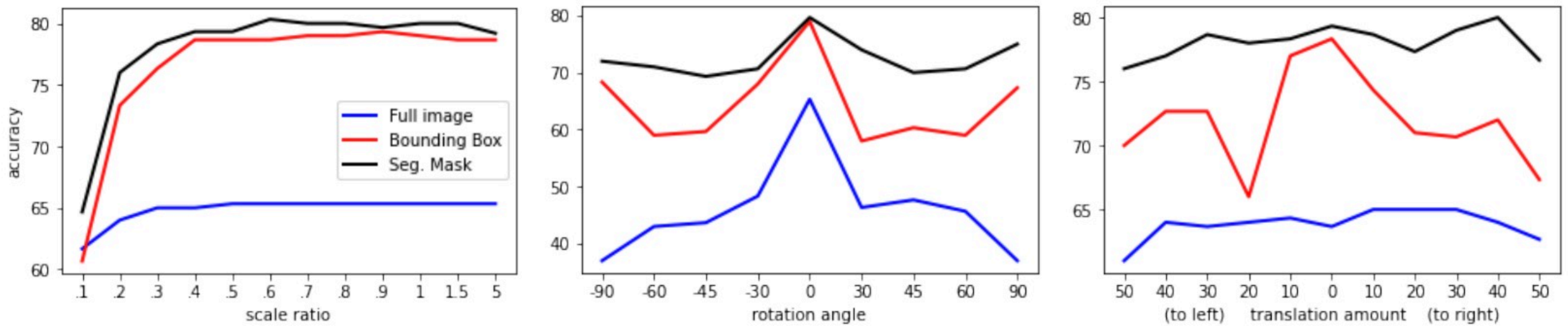
- We trained three ResNet-18 models (finetuned on ImageNet), one per each input type:
    - 1) full image,
    - 2) bounding box,
    - 3) segmented object (placed in a dark background).

- Models were trained on 70 images per category (700 in total) for 10 epochs and were then tested on the remaining 30 images per category.



Model accuracy against adversarial perturbations (left) and noise corruptions (middle). The right panel shows a sample chair image along with its bounding box and segmentation mask.

Robustness Against Geometric Transformations



Performance (top-1) of the ResNet-18 model against geometric transformations.

Qualitative observations

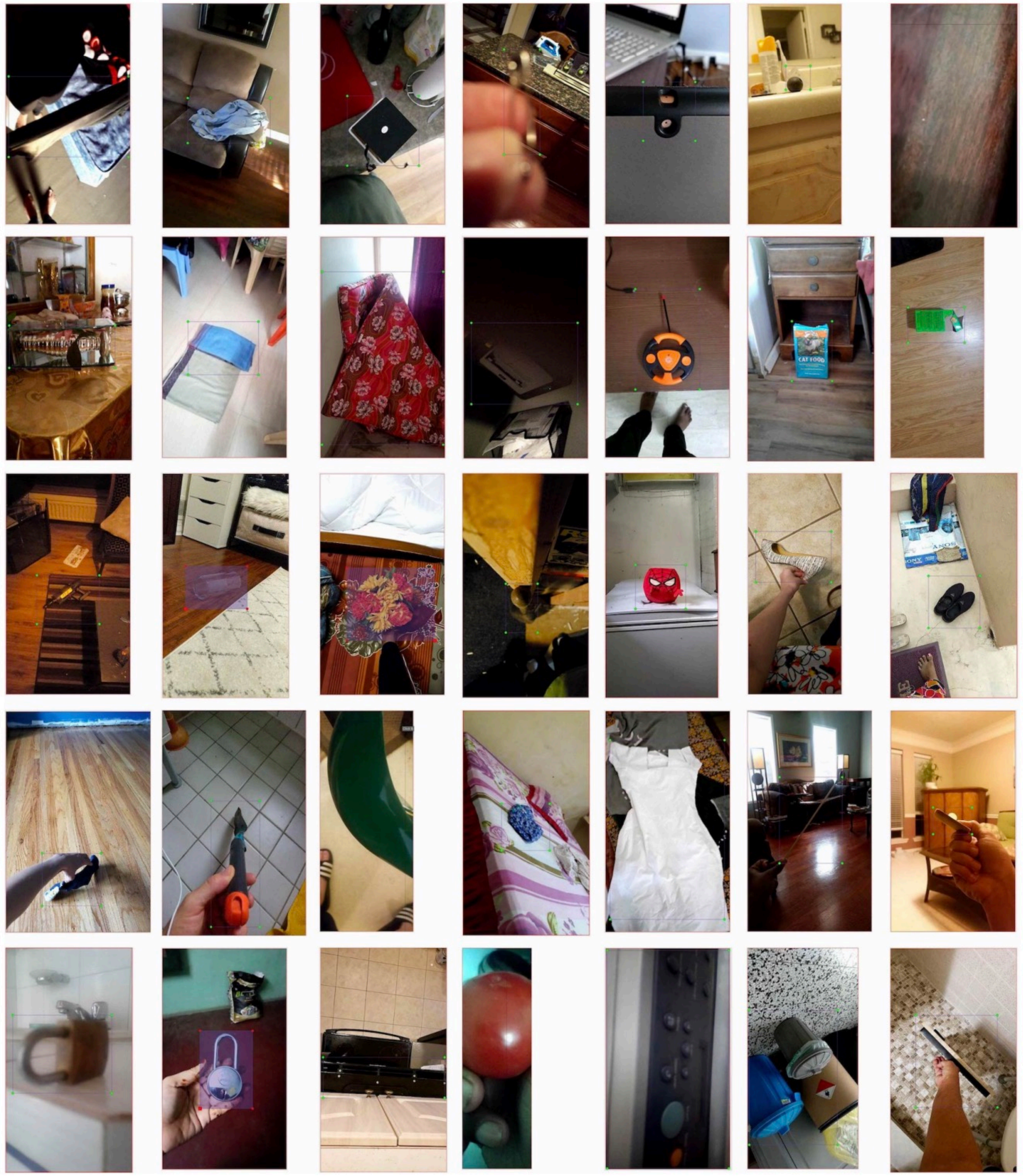
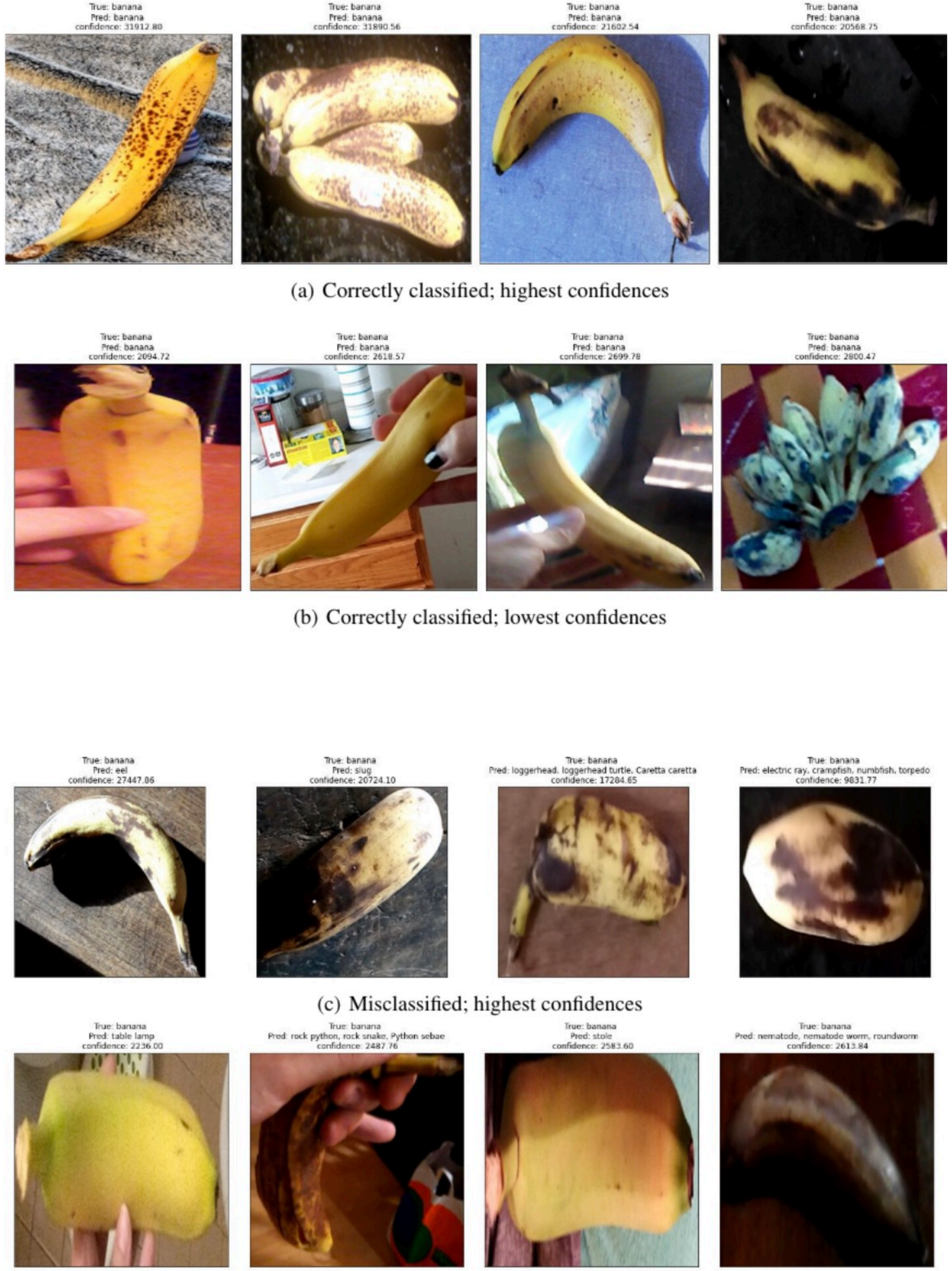


Figure 7: A selection of challenging objects that are hard to be recognized by humans. Can you guess the category of the annotated objects in these images? Keys are as follows:
row 1: *(skirt, skirt, desk lamp, safety pin, still camera, spatula, tray)*,
row 2: *(vase, pillow, sleeping bag, printer, remote control, pet food container, detergent)*,
row 3: *(vacuum cleaner, vase, vase, shovel, stuffed animal, sandal, sandal)*,
row 4: *(sock, shovel, shovel, skirt, skirt, match, spatula)*,
row 5: *(padlock, padlock, microwave, orange, printer, trash bin, tray)*



(a) Correctly classified; highest confidences

(b) Correctly classified; lowest confidences

(c) Misclassified; highest confidences

(d) Misclassified; lowest confidences

Summary & Conclusion

- Significant progress has been made in visual recognition (e.g., object recognition, detection, segmentation, …). Despite high performance, models generalize poorly out of their comfort zone (out of distribution generalization).
- ImageNet dataset served us well and was influential in driving the field forward. With that out of the picture, it is hard to gauge the progress. What next?!
- We analyzed models on a challenging real world dataset and found that object recognition models perform significantly well on bounding boxes rather than the entire scene.
- We also found that limiting the object area as much as possible improves accuracy and robustness. This needs to be verified also in larger scale.
- The role of context in computer vision tasks is still not clear
- **Proposed new task: recognizing objects using their context (i.e, using detection datasets to train recognition models)**
- Code and annotation data is publicly available

KEEP
CALM
AND
STAY
SAFE