Weakly Supervised Object Localization with Progressive Domain Adaptation Supplementary Material

Dong Li[1], Jia-Bin Huang[2], Yali Li[1], Shengjin Wang[1]*, and Ming-Hsuan Yang[3]
[1]Tsinghua University, [2]University of Illinois, Urbana-Champaign, [3]University of California, Merced

1. Overview

In this supplementary material, we present three additional results to complement the paper. First, we report detailed quantitative evaluation on the PASCAL VOC and ILSVRC object detection datasets. Second, we show additional qualitative detection results on the VOC 2007 dataset. Third, we analyze the errors of three variants of the proposed approach and show relative contributions from each component.

2. Quantitative Evaluation

We show in Table 1, 2, and 3 the detection average precision (AP) performance of our method on the PASCAL VOC 2007, 2010 and 2012 datasets, respectively. In general, our algorithm achieves better localization performance for animal and vehicle than that for furniture classes. It is difficult to detect indoor objects in a weakly supervised manner due to large appearance variations and background clutters. Using the deeper model, the VGGNet, as our base CNN model, we achieve better performance than that from the AlexNet.

We also present the precision-recall curves for each category by our OM+MIL+FT-VGGNet method on the VOC 2007 *test* set in Figure 1, the VOC 2010 *val* set in Figure 2, and the VOC 2012 *val* set in Figure 3.

Table 4 and 5 show the per-class detection average precision performance of our method on the ILSVRC 2013 detection val_2 set. Similarly, we achieve better localization performance using the deeper network.

3. Sample detections

We show more sample detection results on th PASCAL VOC 2007 *test* set in Figure 4, 5, 6. All the detection results are obtained by our OM+MIL+FT-VGGNet method. Our algorithm is able to detect objects under different scales, lighting conditions and partial occlusions.

*Corresponding author

4. Error analysis

In Figure 8, we apply the detector error analysis tool from Hoiem et al. [1] to analyze errors of our detector on the VOC 2007 dataset. Comparing the first column (OM+MIL) with the third column (OM+MIL+FT), we find confusion with similar objects is significantly reduced by the detection adaptation step (FT), particularly for furniture classes. Fine-tuning the network for object-level detection helps learn discriminative appearance model for object categories. Comparing the second column (MIL+FT) with the third column (OM+MIL+FT), confusion with background (particularly for vehicle classes) and other objects (particularly for furniture classes) is reduced with our class-specific proposal mining (OM) step. With classification adaptation, the OM step helps remove a substantial amount of noise in the initial noisy class-independent proposal collection and mines out class-specific object candidates with high-precision.

From the error analysis plots, we note that the majority of errors comes from *inaccurate localization*. We show sample detection errors in Figure 7. Typical errors with imprecise localization include detecting a bicycle wheel, a bird body, or partial bus and car. Sometimes the detector gets confused with background or similar objects. The last two rows of Figure 7 show the detection errors due to confusion with background and similar objects, respectively. For example, we detect plant in lake and claim to detect potted plants. In the first image on the last row, we incorrectly detect a chair as a sofa.

References

[1] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *ECCV*, 2012. 1

Table 1. Quantitative evaluation in terms of detection average precision (AP) on the PASCAL VOC 2007 *test* set.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
OM + MIL + FT-AlexNet	49.7	33.6	30.8	19.9	13	40.5	54.3	37.4	14.8	39.8	9.4	28.8	38.1	49.8	14.5	24	27.1	12.1	42.3	39.7	31.0
OM + MIL + FT-VGGNet	54.5	47.4	41.3	20.8	17.7	51.9	63.5	46.1	21.8	57.1	22.1	34.4	50.5	61.8	16.2	29.9	40.7	15.9	55.3	40.2	39.5

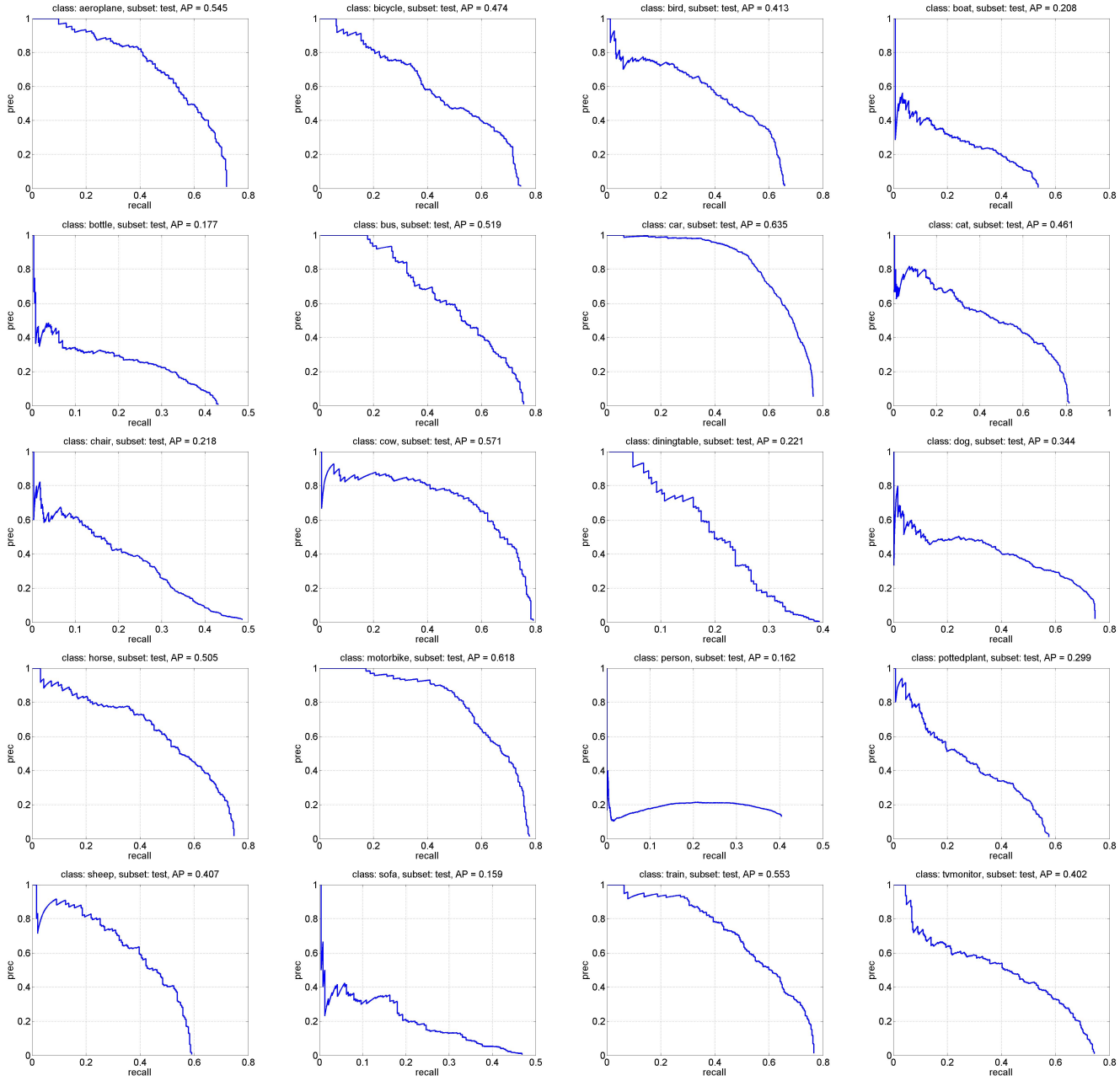

Figure 1. Precision-recall curves for each category on the PASCAL VOC 2007 *test* set.

Table 2. Quantitative evaluation in terms of detection average precision (AP) on the PASCAL VOC 2010 *val* set.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
OM + MIL + FT-AlexNet	37	26.7	17.3	7.1	8.7	55.3	27.4	34.8	6.2	18.7	4.2	20.7	28.4	39.6	8.9	11.3	22.6	6.4	23.3	22.7	21.4
OM + MIL + FT-VGGNet	46.7	50.9	31.9	9.2	13.2	54.6	35.4	47.1	7.5	24.2	6.2	32.6	46.3	56.8	11.6	21.2	32.6	12	45.5	28.9	30.7

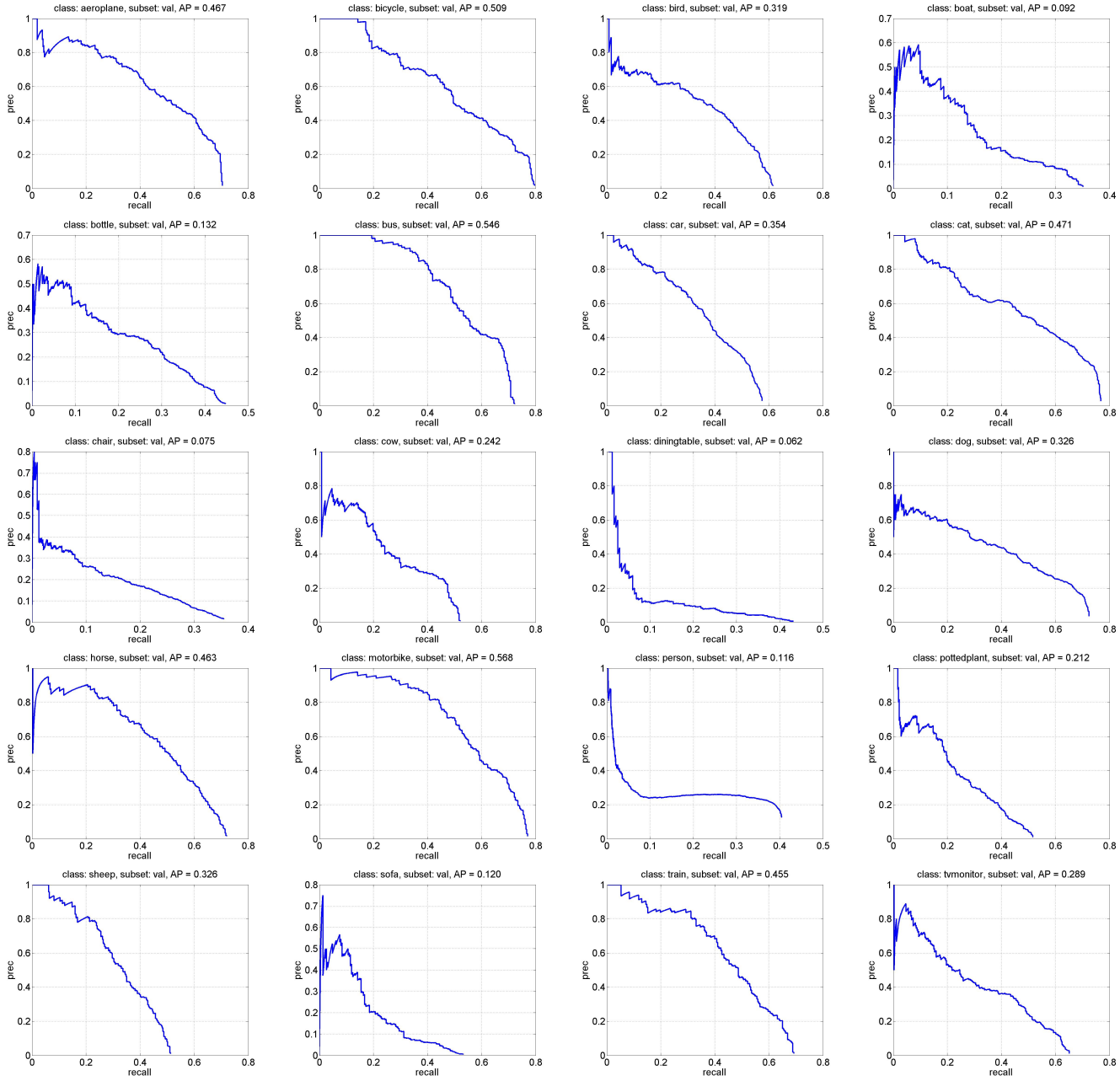

Figure 2. Precision-recall curves for each category on the PASCAL VOC 2010 *val* set.

Table 3. Quantitative evaluation in terms of detection average precision (AP) on the PASCAL VOC 2012 *val* set.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
OM + MIL + FT-AlexNet	33.7	27.8	19.9	5.7	6.5	49.6	28.1	37.3	7.6	17.4	2.8	23.3	33.7	40.4	8.9	13.7	22.3	5.5	36.4	26.7	22.4
OM + MIL + FT-VGGNet	42.2	27.8	32.7	4.2	13.7	52.1	35.8	48.3	11.8	31.7	4.9	30.4	45.3	51.8	11.5	13.4	33.5	7.2	45.6	38.4	29.1

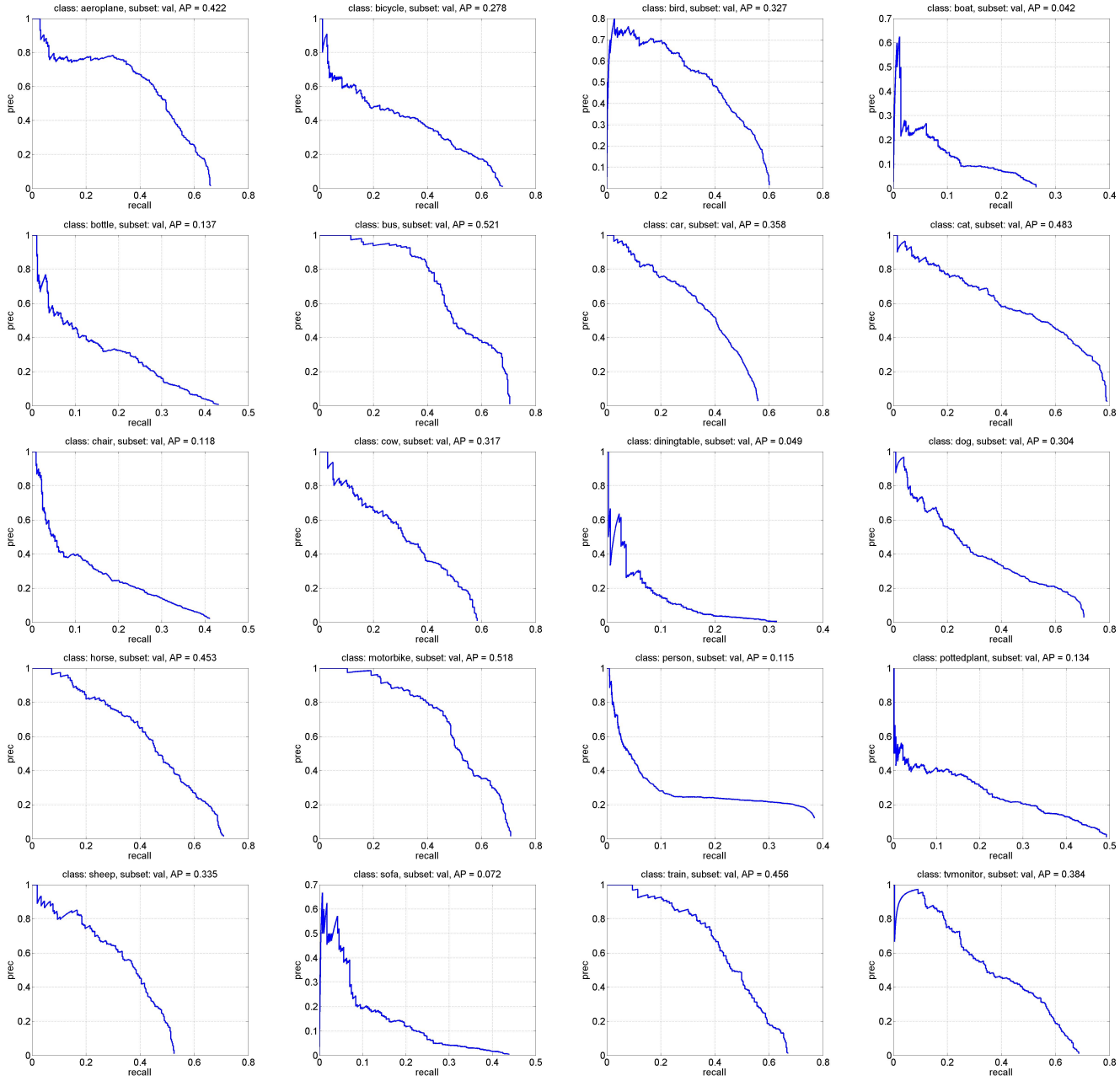

Figure 3. Precision-recall curves for each category on the PASCAL VOC 2012 *val* set.

Table 4. Per-class detection average precision (AP) on the ILSVRC2013 detection val_2 set by our OM+MIL+FT-AlexNet method.

class	AP	class	AP	class	AP	class	AP	class	AP
accordion	14.0	centipede	12.5	hair spray	4.4	pencil box	4.2	snowplow	21.4
airplane	14.9	chain saw	1.1	hamburger	15.3	pencil sharpener	0.7	soap dispenser	0.1
ant	18.7	chair	3.9	hammer	0.4	perfume	12.6	soccer ball	14.3
antelope	9.3	chime	12.7	hamster	26.9	person	0.5	sofa	2.7
apple	9.5	cocktail shaker	6.9	harmonica	0.4	piano	13.2	spatula	0.2
armadillo	38.0	coffee maker	1.5	harp	13.4	pineapple	9.5	squirrel	14.2
artichoke	3.1	computer keyboard	4.9	hat with a wide brim	1.6	ping-pong ball	0.0	starfish	9.5
axe	0.2	computer mouse	0.0	head cabbage	1.0	pitcher	5.1	stethoscope	0.1
baby bed	10.6	corkscrew	1.0	helmet	0.3	pizza	3.9	stove	0.6
backpack	1.0	cream	2.0	hippopotamus	13.3	plastic bag	0.1	strainer	1.5
bagel	3.1	croquet ball	0.0	horizontal bar	0.5	plate rack	0.2	strawberry	5.1
balance beam	0.1	crutch	0.1	horse	14.4	pomegranate	1.1	stretcher	0.4
banana	5.2	cucumber	2.4	hotdog	7.0	popsicle	1.8	sunglasses	1.6
band aid	0.6	cup or mug	12.7	iPod	23.6	porcupine	17.2	swimming trunks	0.0
banjo	7.6	diaper	0.0	isopod	9.2	power drill	1.0	swine	27.1
baseball	12.3	digital clock	2.5	jellyfish	3.2	pretzel	4.2	syringe	0.0
basketball	0.0	dishwasher	0.1	koala bear	13.1	printer	3.4	table	0.8
bathing cap	1.2	dog	24.0	ladle	0.0	puck	0.0	tape player	8.8
beaker	6.5	domestic cat	8.8	ladybug	16.6	punching bag	3.0	tennis ball	12.3
bear	33.3	dragonfly	17.0	lamp	1.1	purse	4.6	tick	29.8
bee	12.4	drum	0.5	laptop	1.4	rabbit	35.2	tie	0.1
bell pepper	3.7	dumbbell	2.0	lemon	10.2	racket	0.0	tiger	14.5
bench	1.2	electric fan	0.0	lion	3.4	ray	6.9	toaster	22.9
bicycle	12.1	elephant	23.2	lipstick	4.6	red panda	12.0	traffic light	0.5
binder	1.0	face powder	5.8	lizard	4.5	refrigerator	1.9	train	21.1
bird	28.4	fig	1.9	lobster	4.2	remote control	4.7	trombone	0.2
bookshelf	2.4	filing cabinet	4.9	maillot	0.2	rubber eraser	0.1	trumpet	3.3
bow tie	0.0	flower pot	0.3	maraca	11.3	rugby ball	0.1	turtle	25.3
bow	0.6	flute	0.2	microphone	0.0	ruler	0.4	tv or monitor	8.1
bowl	4.6	fox	28.1	microwave	0.8	salt or pepper shaker	6.6	unicycle	0.2
brassiere	0.2	french horn	0.5	milk can	11.8	saxophone	2.7	vacuum	2.5
burrito	3.3	frog	11.2	miniskirt	0.0	scorpion	23.7	violin	1.6
bus	21.0	frying pan	0.4	monkey	19.2	screwdriver	0.1	volleyball	0.0
butterfly	62.3	giant panda	41.1	motorcycle	20.5	seal	1.6	waffle iron	1.6
camel	5.3	goldfish	8.8	mushroom	11.4	sheep	10.9	washer	7.7
can opener	2.5	golf ball	17.4	nail	0.0	ski	0.2	water bottle	2.0
car	14.9	golfcart	54.7	neck brace	0.1	skunk	8.5	watercraft	6.0
cart	23.3	guacamole	12.0	oboe	9.8	snail	11.3	whale	0.0
cattle	9.8	guitar	6.4	orange	3.8	snake	5.1	wine bottle	1.5
cello	8.8	hair dryer	1.6	otter	2.2	snowmobile	2.5	zebra	13.4

Table 5. Per-class detection average precision (AP) on the ILSVRC2013 detection *val*$_2$ set by our OM+MIL+FT-VGGNet method.

class	AP	class	AP	class	AP	class	AP	class	AP
accordion	15.9	centipede	14.6	hair spray	1.8	pencil box	3.4	snowplow	18.0
airplane	24.0	chain saw	4.8	hamburger	12.9	pencil sharpener	0.5	soap dispenser	1.1
ant	24.6	chair	4.7	hammer	0.5	perfume	19.8	soccer ball	21.7
antelope	13.7	chime	15.3	hamster	43.0	person	0.6	sofa	5.2
apple	8.1	cocktail shaker	7.3	harmonica	4.5	piano	6.0	spatula	0.1
armadillo	50.5	coffee maker	6.0	harp	28.9	pineapple	16.8	squirrel	23.8
artichoke	4.1	computer keyboard	9.8	hat with a wide brim	5.1	ping-pong ball	0.6	starfish	18.8
axe	1.7	computer mouse	0.0	head cabbage	4.9	pitcher	11.9	stethoscope	2.2
baby bed	19.3	corkscrew	10.2	helmet	0.5	pizza	6.5	stove	2.5
backpack	0.7	cream	4.4	hippopotamus	6.7	plastic bag	1.7	strainer	4.9
bagel	7.4	croquet ball	0.1	horizontal bar	0.4	plate rack	3.4	strawberry	8.4
balance beam	0.1	crutch	1.1	horse	19.8	pomegranate	4.4	stretcher	0.7
banana	6.9	cucumber	4.7	hotdog	11.0	popsicle	1.6	sunglasses	2.3
band aid	3.2	cup or mug	17.8	iPod	31.2	porcupine	40.9	swimming trunks	0.0
banjo	6.7	diaper	0.8	isopod	24.2	power drill	2.1	swine	39.6
baseball	18.1	digital clock	9.0	jellyfish	1.0	pretzel	2.0	syringe	1.3
basketball	0.0	dishwasher	0.6	koala bear	21.1	printer	8.7	table	0.9
bathing cap	0.8	dog	40.2	ladle	0.6	puck	0.0	tape player	9.6
beaker	4.4	domestic cat	22.1	ladybug	20.6	punching bag	8.4	tennis ball	17.5
bear	45.8	dragonfly	26.0	lamp	1.5	purse	8.1	tick	7.3
bee	11.4	drum	0.3	laptop	4.7	rabbit	47.1	tie	0.1
bell pepper	9.6	dumbbell	4.4	lemon	11.7	racket	0.5	tiger	27.3
bench	1.8	electric fan	0.3	lion	17.6	ray	12.6	toaster	23.2
bicycle	13.6	elephant	33.6	lipstick	1.8	red panda	11.0	traffic light	0.6
binder	2.2	face powder	5.5	lizard	19.1	refrigerator	15.0	train	19.0
bird	40.8	fig	4.4	lobster	11.3	remote control	11.3	trombone	2.0
bookshelf	1.3	filing cabinet	6.2	maillot	0.4	rubber eraser	0.1	trumpet	5.5
bow tie	0.0	flower pot	0.8	maraca	11.1	rugby ball	0.1	turtle	35.6
bow	3.3	flute	0.0	microphone	0.0	ruler	0.9	tv or monitor	6.6
bowl	6.9	fox	35.0	microwave	3.4	salt or pepper shaker	2.8	unicycle	0.4
brassiere	1.6	french horn	2.0	milk can	22.3	saxophone	11.5	vacuum	10.8
burrito	4.2	frog	27.0	miniskirt	0.1	scorpion	28.1	violin	4.8
bus	25.5	frying pan	2.4	monkey	29.8	screwdriver	0.1	volleyball	0.0
butterfly	64.9	giant panda	49.1	motorcycle	21.0	seal	5.8	waffle iron	3.9
camel	9.2	goldfish	7.9	mushroom	13.2	sheep	20.8	washer	3.7
can opener	8.1	golf ball	22.2	nail	0.0	ski	0.1	water bottle	2.9
car	19.4	golfcart	55.0	neck brace	0.0	skunk	8.6	watercraft	12.0
cart	18.5	guacamole	14.4	oboe	1.4	snail	13.3	whale	0.0
cattle	17.0	guitar	11.2	orange	3.2	snake	13.7	wine bottle	1.8
cello	6.1	hair dryer	4.4	otter	5.0	snowmobile	6.0	zebra	29.5

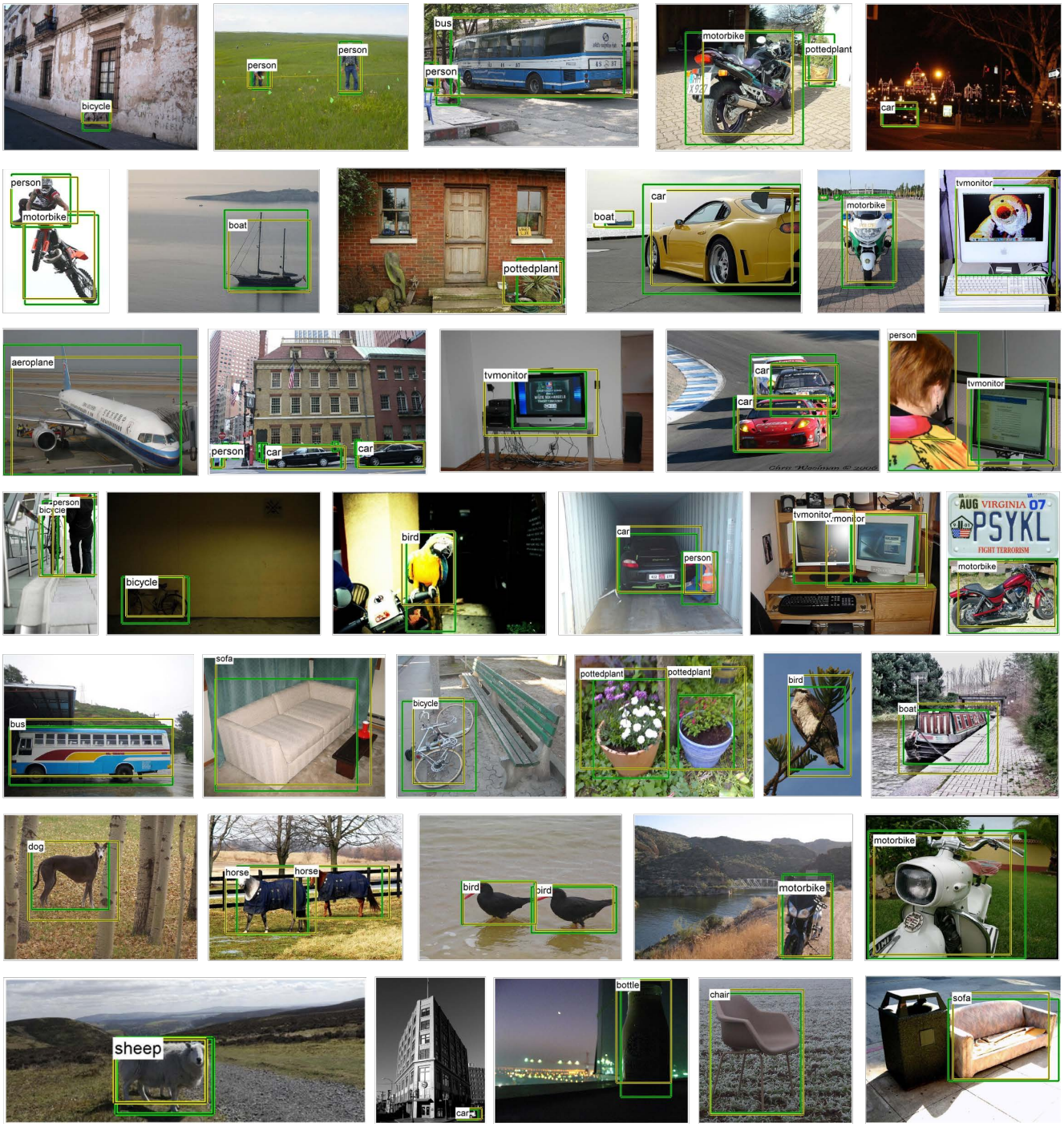

Figure 4. Sample detection results on the PASCAL VOC 2007 *test* set. Green boxes indicate ground-truth instance annotation. Yellow boxes indicate correction detections (with IoU $\geq$ 0.5). For all the testing results, we set threshold of detection as 0.8 and use NMS to remove duplicate detections.

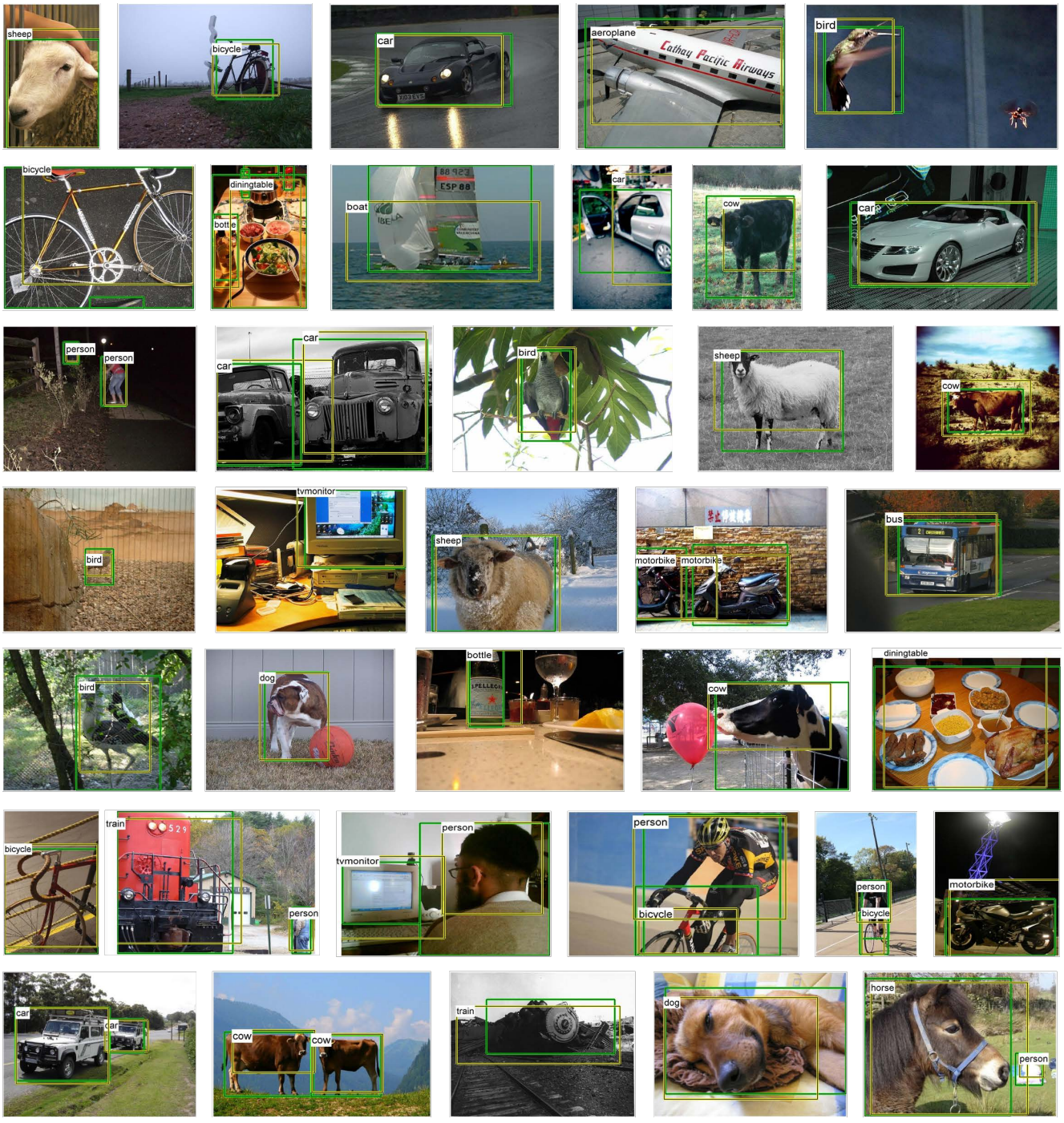

Figure 5. Sample detection results on the PASCAL VOC 2007 *test* set. Green boxes indicate ground-truth instance annotation. Yellow boxes indicate correction detections (with IoU $\geq$ 0.5). For all the testing results, we set threshold of detection as 0.8 and use NMS to remove duplicate detections.

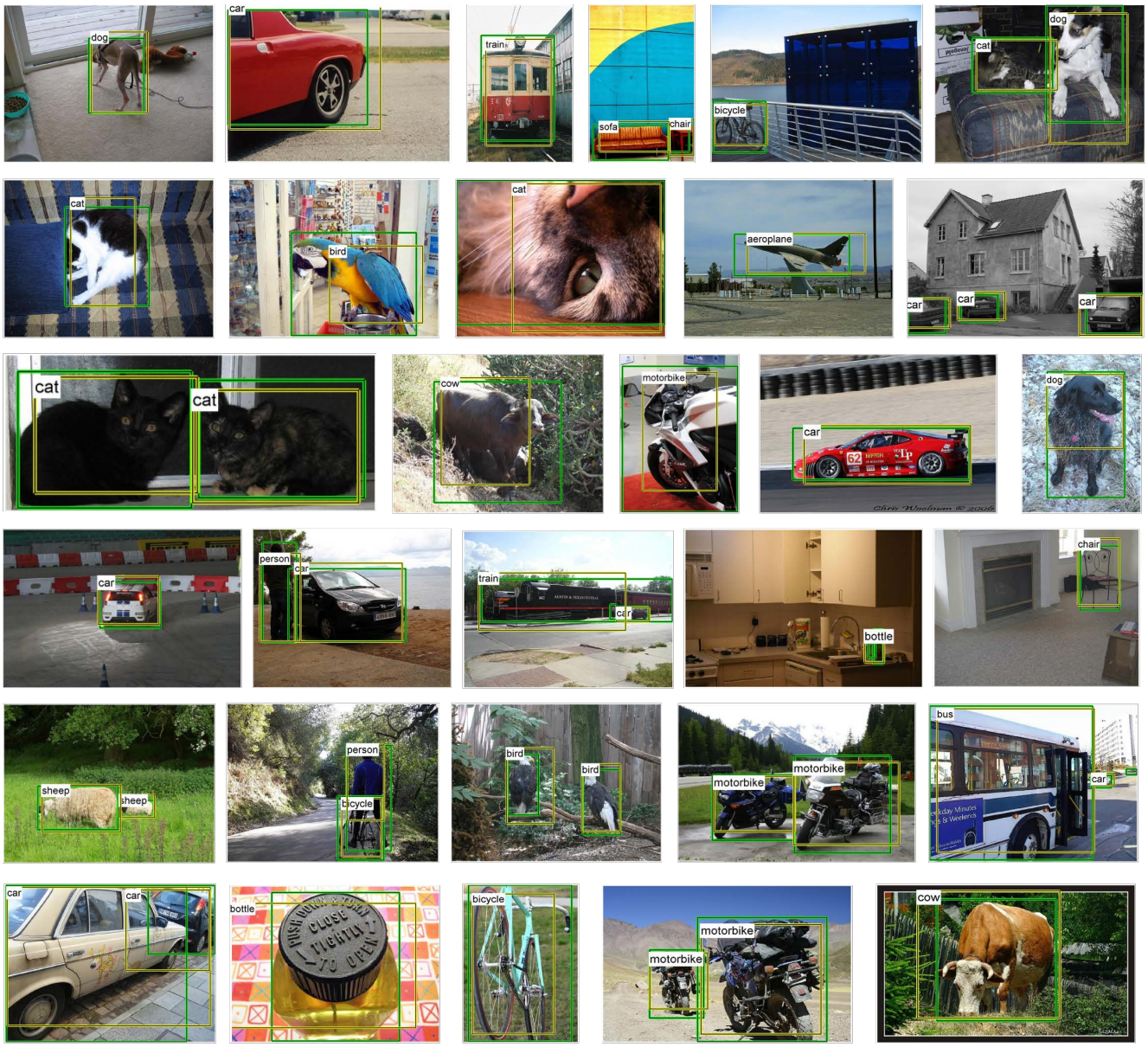

Figure 6. Sample detection results on the PASCAL VOC 2007 *test* set. Green boxes indicate ground-truth instance annotation. Yellow boxes indicate correction detections (with IoU $\geq$ 0.5). For all the testing results, we set threshold of detection as 0.8 and use NMS to remove duplicate detections.

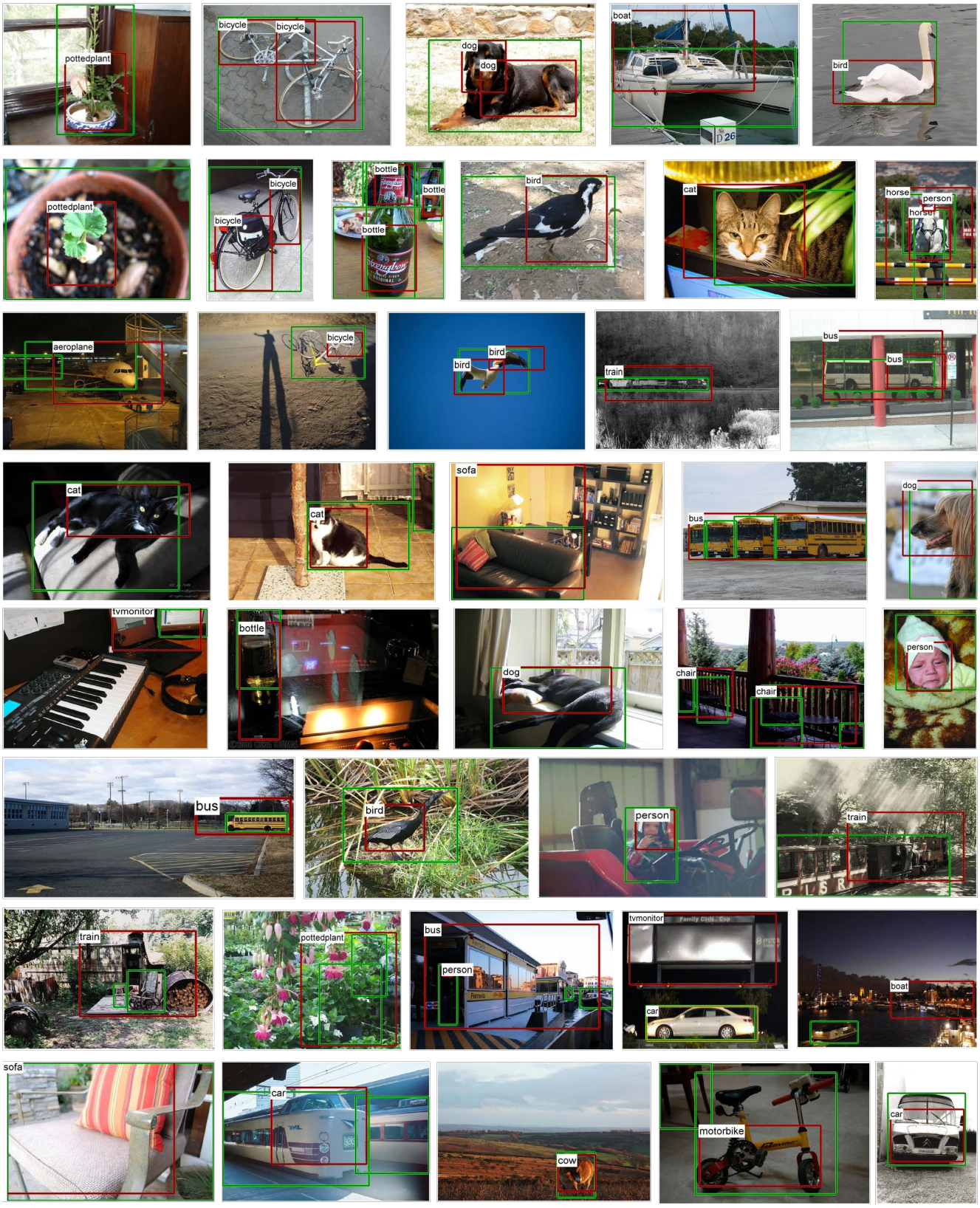

Figure 7. Additional sample results of detection errors. Green boxes indicate ground-truth instance annotation. Red boxes indicate false positives.

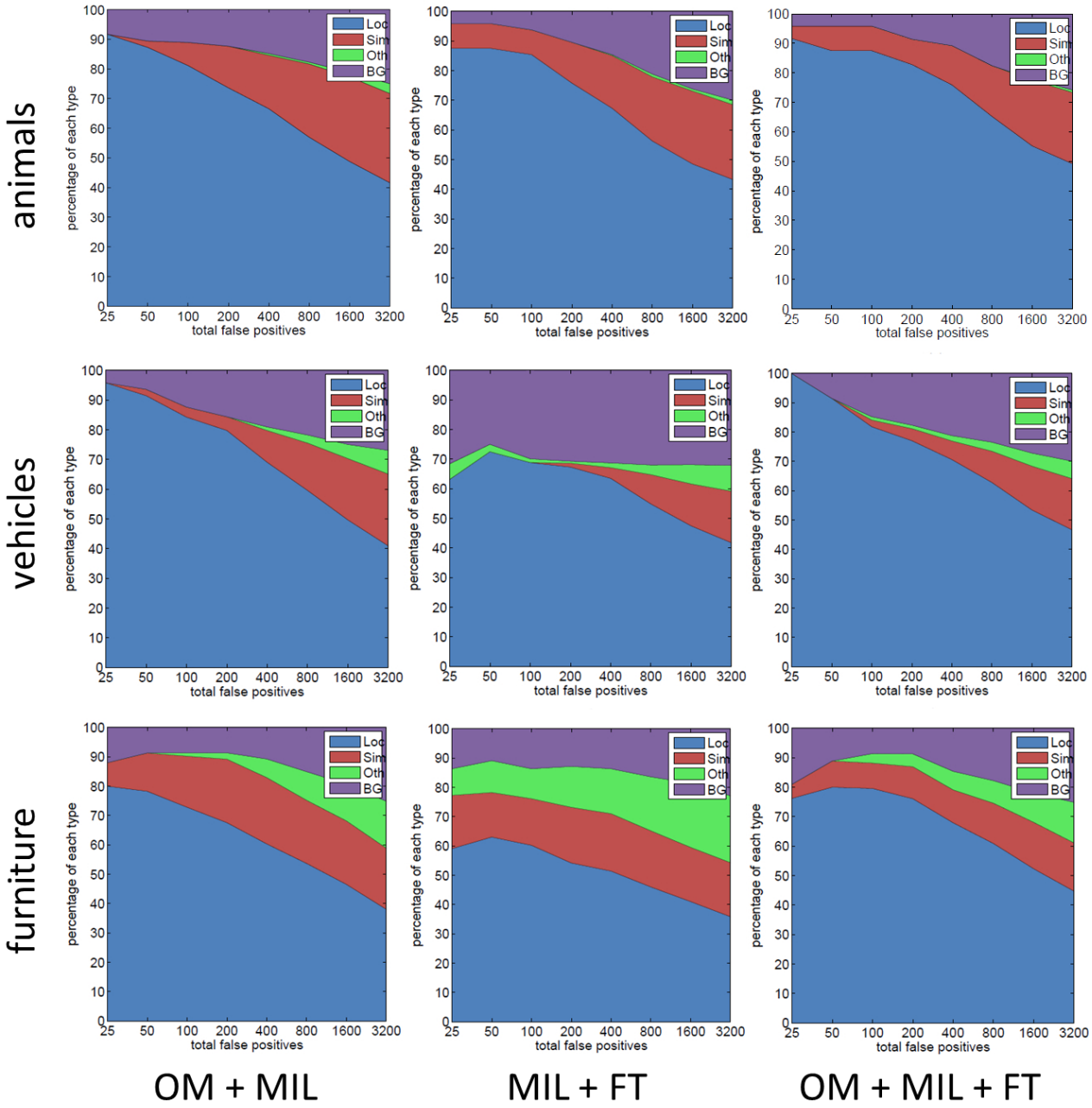

Figure 8. Detector error analysis. The detection errors are categorized into four types: false positives due to poor localization (Loc), confusion with similar objects (Sim), confusion with other VOC objects (Oth), and confusion with background (BG). The stacked area plots show fraction of FP of each type as the total number of FP increase. We take examples of animal, vehicle and furniture classes. The results of "MIL+FT" and "OM+MIL+FT" are obtained using the VGGNet.