Supplementary Materials

Discriminative Bimodal Networks for Visual Localization and Detection with Natural Language Queries

Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, Honglak Lee
University of Michigan, Ann Arbor, MI, USA
{yutingzh, yuanluya, guoyijie, zhiyuan, huangian, honglak}@umich.edu

Contents

A. CNN architecture for the linguistic pathway 1
B. Formalized comparison with conditional generative models 1
C. Model optimization 2
D. Discussion on recall and precision for localization 2
E. More quantitative results 2
E.1 . Pretraining on different datasets 2
E.2 . Random and oracle localization performance 3
E.3 . Localization using constrained queries 3
E.4 . Ablative study on the text similarity threshold 4
F . More qualitative comparison for localization 4
F.1 . More qualitative comparison with DenseCap 5
F.2 . More qualitative comparison with SCRC 6
G. Qualitative Comparison for Detection 7
G.1. Random detection results with known number of ground truths 8
G.2. Random detection results with phrase-dependent thresholds 11
G.3. Failure cases for detection with phrase-dependent thresholds 20
H. Precision-recall curves 23
H.1. Phrase-independent precision-recall curves 23
H.2. Phrase-dependent precision-recall curves 24

A. CNN architecture for the linguistic pathway

We summarize the CNN architecture used for the linguistic pathway in Table 5.

Layer ID	Type	Kernel size	Output channels	Pooling size	Output length	Activation
0	input	n/a	74	none	256	none
1	convolution	7	256	2	128	LReLU (leakage $= 0.1$)
2	convolution	7	256	none	128	LReLU (leakage $= 0.1$)
3	convolution	3	256	none	128	LReLU (leakage $= 0.1$)
4	convolution	3	256	2	64	LReLU (leakage $= 0.1$)
5	convolution	3	512	none	64	LReLU (leakage $= 0.1$)
6	convolution	3	512	2	32	LReLU (leakage $= 0.1$)
7	inner-product	n/a	2048	n/a	n/a	LReLU (leakage $= 0.1$)
8	inner-product	n/a	2048	n/a	n/a	LReLU (leakage $= 0.1$)

Table 5: CNN architecture for the linguistic pathway.

B. Formalized comparison with conditional generative models

In contrast to our discriminative framework, which fits $p(l|x, r, t)$, existing methods on natural-language visual localization [21, 23, 38] use the conditional caption generation model, where $f(x, t, r; \Theta)$ resembles $p(t|x, r)$. In [21, 23], the models are trained by maximizing $p(t|x, r)$. In [38], the model is trained instead by maximizing $p(r|x, t)$. However, it still resembles $p(t|x, r)$, and $p(r|x, t)$ is calculated via Bayes' theorem.

Since the space of the natural language is intractable, accurately modeling $p(t|x, r)$ is extremely difficult. Even considering only the plausible text phrases for r on x, the modes of $p(t|x, r)$ are still hard to be properly lifted and balanced due to the lack of enough training samples to cover all valid descriptions. The generative modeling for text phrases may fundamentally limit the discriminative power of the existing model.

In contrast, our model takes both r and t as conditional variables. The conditional distribution on l is much easier to model due to the small binary label space, and it also naturally admits discriminative training. The power of deep distributed representations can also be leveraged for generalizing textual representations to less frequent phrases.

C. Model optimization

The training objective is optimized by back-propagation [32] using the mini-batch stochastic gradient descent (SGD) with momentum 0.9. We use the basic SGD for the visual pathway and Adam [28] for the rest of the network.

We use EdgeBox [62] to propose 1000 boxes per image (in addition to the boxes annotated with text phrases) during training. For each image per iteration, we always include the top 50 proposed boxes in the SGD, and randomly sample another 50 out of the remaining 950 box proposals for diversity and efficiency.

To calculate L_i^{rest} exactly, we need to extract features from all text phrases (>2.8M in Visual Genome) in the training set and combine them with almost every image regions in the mini-batch, which is impractical. Following the stochastic optimization framework, we randomly sample a few text phrases according to their frequencies of occurrence in the training set. This stochastic optimization procedure is consistent with (13).

In each iteration, we sample 2 images when using the 16-layer VGGNet and 1 image when using ResNet-101 on a single Titan X. The representations for each unique phrase and each unique image region is computed once per iteration. We partition a DBNet into sub-networks for the visual and textual pathways, and for the discriminative pathway. The batch size for those sub-networks are different and determined by inputs, e.g., the numbers of text phrases, bounding boxes, and effective region-text pairs. When using 2 images per iteration, the batch size for the discriminative pathway is $\sim$10K, where we feed all effective region-text pairs, as defined in (9) , to the discriminative pathway. The large batch size is needed for efficient and stable optimization. Our Caffe [22] and MATLAB based implementation supports dynamic and arbitrarily large batch sizes for sub-networks. The initial learning rates when using different visual pathways are summarized in Table 6.

Sub-networks \ Models		16-layer VGGNet	ResNet-101
Visual pathway	Before RoI-pooling	10^{-3}	10^{-3}
	After RoI-pooling	10^{-3}	10^{-4}
Remainder		10^{-4}	10^{-5}

Table 6: Learning rates for DBNet training

We trained the VGG-based DBNet for approximately 10 days (3–4 days without finetuning the visual network, 4–5 days for the whole network, and 1–2 days with the decreased learning rate). DenseCap could get converged in $\sim$4 days, but further training did not improve the results. Given DBNet's much higher accuracy, the extra training time was worthwhile.

D. Discussion on recall and precision for localization

Table 1, 2, and 4 report the recall for the localization tasks, where each text phrase is localized with the bounding box of the highest score. Given an IoU threshold, the localized bounding box is either correct or not. As no decision threshold exists in this setting, we can calculate only the *accuracy*, but not a precision-recall curve. Following the convention in DenseCap and SCRC, we call this accuracy the "(rank-1) *recall*", since it reflects if any ground-truth region can be recalled by the top-scored box. In Figure 3, assuming one ground-truth region per image (i.e., ordinary localization settings), we have $\text{precision} = \text{recall}/\text{rank}$. Note that rank-1 precision is the same as rank-1 recall.

E. More quantitative results

We provide more quantitative analysis in this section, including the impact of pretraining on other datasets, random and upper-bound localization performance, localization with controlled queries, and an ablative study on the text similarity threshold for determining the ambiguous text phrase set.

E.1. Pretraining on different datasets

We trained DBNet and DenseCap using various pretrained visual networks. In particular, we used the 16-layer VGGNet in two settings: 1) pretrained on ImageNet ILSVRC 2012 for image classification (VGGNet-CLS) [8] and 2) further pretrained on the PASCAL VOC [10] for object detection using faster R-CNN [46]. We compared DBNet and DenseCap trained with these two pretrained networks and tested them with two different region proposal methods (i.e., DenseCap RPN and EdgeBox). As shown in Table 7, VOC pretraining was beneficial for DBNet, but it was not beneficial for DenseCap. Thus, we used the ImageNet pretrained VGGNet for DenseCap in the main paper.

Region	Localization	Accuracy / % for IoU@							Median	Mean
proposal	model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	IoU	IoU
DC-RPN 500	DenseCap (VGGNet-CLS)	52.5	38.9	27.0	17.1	9.5	4.3	1.5	0.117	0.184
	DenseCap (VGGNet-DET)	49.4	36.9	26.0	16.7	9.3	4.3	1.5	0.096	0.176
	DBNet (VGGNet-CLS)	**57.7**	**46.9**	37.0	27.9	19.5	11.7	5.6	0.169	0.242
	DBNet (VGGNet-DET)	57.4	**46.9**	**37.8**	**29.4**	**21.3**	**13.6**	**7.0**	**0.168**	**0.250**
EdgeBox 500	DenseCap (VGGNet-CLS)	48.8	36.2	25.7	16.9	10.1	5.4	2.4	0.092	0.178
	DenseCap (VGGNet-DET)	46.6	34.8	24.9	16.6	10.0	5.2	2.2	0.076	0.171
	DBNet (VGGNet-CLS)	54.3	45.0	36.6	28.8	21.3	14.4	8.2	0.144	0.245
	DBNet (VGGNet-DET)	**54.8**	**45.9**	**38.3**	**30.9**	**23.7**	**16.6**	**9.9**	**0.152**	**0.258**

Table 7: Localization performance for DBNet and DenseCap with different pretrained models on Visual Genome. VGGNet-CLS: the 16-layer VGGNet pretrained on ImageNet ILSVRC 2012 dataset. VGGNet-DET: the 16-layer VGGNet further pretrained on PASCAL VOC07+12 trainval set.

E.2. Random and oracle localization performance

Given proposed image regions, we performed localization for text phrases with random guessing and the oracle detector. For random guessing, we randomly chose a proposed region and took it as the localization results. For more accurate evaluation, we averaged the results over all possible cases (i.e., enumerating over all proposed boxes). For the oracle detector, it always picked up the proposed region that had the largest overlap with a ground truth region, providing the performance upper bound due to the limitation of the region proposal method, as in [61].

As shown in Table 8, the trained models (DBNet, SCRC, DenseCap) significantly outperformed random guessing, which suggests that promising models can be developed using deep neural networks. However, the the performance of DBNet had a large gap with the oracle detector, which indicates that more advanced methods need to be developed in the further to better address the natural language visual localization problem.

Model	Recall / % for IoU@							Median	Mean
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	IoU	IoU
Random	19.0	10.0	5.2	2.6	1.2	0.5	0.2	0.041	0.056
DenseCap	48.8	36.2	25.7	16.9	10.1	5.4	2.4	0.092	0.178
SCRC	52.0	39.1	27.8	18.4	11.0	5.8	2.5	0.115	0.189
DBNet	**54.8**	**45.9**	**38.3**	**30.9**	**23.7**	**16.6**	**9.9**	**0.152**	**0.258**
Oracle	94.0	87.3	80.4	73.1	65.1	55.8	42.4	0.650	0.572

Table 8: Single-image object localization accuracy on the Visual Genome dataset for random guess, oracle detector, and trained models. EdgeBox is used to propose 500 regions per image. *Random*: a proposed region is randomly chosen as the localization for a text phrase and the performance is averaged over all possibilities; *Oracle*: the proposed region that has the largest overlap with the ground box(es) is taken as the localization for a text phrase.

E.3. Localization using constrained queries

Pairwise relationships describe a particular type of visual entities, i.e., two objects interacting with each other in a certain way. As the basic building block of more complicated parsing structures, the pairwise relationship is worth evaluating as a special case. The Visual Genome dataset has pairwise object relationship annotations, independent from the text phrase annotations. To fit "object-relationship-object" (Obj-Rel-Obj) triplets into our model, we represented a triplet in a SVO (subject-verb-object) text phrase, and took the bounding box enclosing the two objects as the ground truth region for the SVO phrase. During the training time, we used both the original text phrase annotations and the SVO phrases derived from the relationship annotations to keep sufficient diversity of the text descriptions. During the testing time, we used only the SVO phrases to focus on the localization of pairwise relationships. The training and testing sets of images were the same as in the other experiments.

As reported in Table 1, the localization recall for the IoU threshold at 0.5 was close to 50%. The groups of two objects were easier to localize than general visual entities, since they were more clearly defined and generally context-free. In particular, DBNet's performance (recall and median/mean IoU) for Obj-Rel-Obj was approximately twice as high as that for general text phrases. The above experimental results demonstrate the effectiveness of DBNet for localizing object relationships. The results also demonstrate the complexity of the text quires (e.g., using all human-annotated phrases v.s. obj-rel-obj pairs) as a significant source of failures.

Region proposal	Visual network	Localization model	Recall / % for IoU@							Median IoU	Mean IoU
			0.1	0.2	0.3	0.4	0.5	0.6	0.7		
EdgeBox 500	16-layer VGGNet	DBNet (all phrases)	**54.8**	**45.9**	**38.3**	**30.9**	**23.7**	**16.6**	**9.9**	**0.152**	**0.258**
		DBNet (Obj-Rel-Obj)	**81.8**	**75.1**	**67.3**	**57.8**	**46.8**	**35.4**	**23.1**	**0.471**	**0.448**

Table 9: Single-image object localization accuracy on the Visual Genome dataset. Any text phrase annotated on a test image is taken as a query for that image. "IoU@" denotes the overlapping threshold for determining the recall of ground truth boxes. DC-RPN is the region proposal network from DenseCap.

E.4. Ablative study on the text similarity threshold

As discussed in Section 5.3, removing ambiguous training samples are important. The ambiguous sample pruning depends on 1) overlaps between proposed regions and ground truth regions, and 2) text similarity. While the image region overlaps have been commonly considered in traditional object detection, the text similarity is specific to natural language visual localization and detection.

In Table 10, we reported the localization performance of DBNet under different values of the text similarity threshold τ (defined in Eq. (7)), where we considered a controlled setting with neither text phrases from other images nor the visual pathway finetuning. DBNet achieved the best performance with the default parameter $\tau = 0.3$. Suboptimal τ caused approximately 0.5%–1% decrease in localization recall and 0.01 decrease in median/mean IoU.

Phrases from other images	Finetuning visual pathway	τ	Recall / % for IoU@			Median IoU	Mean IoU
			0.3	0.5	0.7		
No	No	0.1	33.6	20.6	8.6	0.101	0.231
No	No	0.2	33.0	20.2	8.5	0.094	0.227
No	No	0.3	**34.5**	**21.2**	**9.0**	**0.113**	**0.237**
No	No	0.4	33.0	20.2	8.4	0.093	0.227
No	No	0.5	32.8	20.2	8.4	0.091	0.226

Table 10: Ablative study on text similarity threshold τ in Eq. (7).

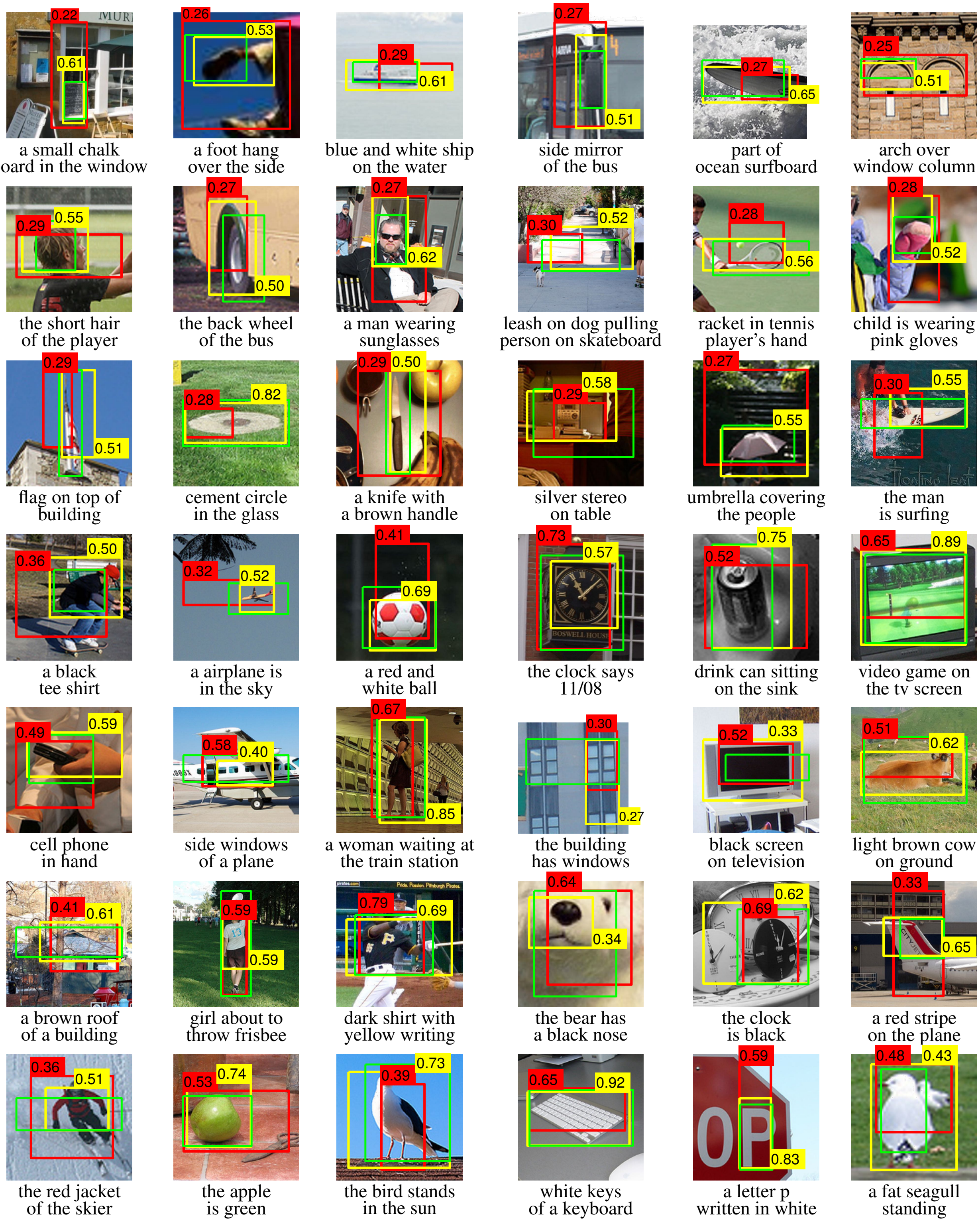
Since the above controlled setting excluded text phrases from the rest of the training set, the localization performance was not too sensitive to the value of τ due to the limited number of phrases. When the text phrases from the whole training set are included in the training loss on a single image, the choice of τ can have a more obvious impact. For example, setting $\tau = 0$ can disable the inclusion of text phrases from other images in any case.

F. More qualitative comparison for localization

More quantitative localization results were shown in this section. We compared DBNet with DenseCap (Figure 6 in Section F.1) and SCRC (Figure 7 in Section F.2), respectively. For each test example, we cropped the image to make the figure focus on the localized region. We used a green box for the ground truth region, a red box for DenseCap/SCRC, and a yellow box for our DBNet.

In the examples that we showed, at least one of the two methods (DBNet and DenseCap/SCRC) can localize the text query to an image region that has IoU > 0.2 overlap with the ground truth region. Besides this constraint, all examples were chosen randomly. While DenseCap and SCRC outperformed DBNet in a few cases, DBNet significantly outperformed those two methods most of the time.

F.1. More qualitative comparison with DenseCap



a small chalk board in the window

a foot hang over the side

blue and white ship on the water

side mirror of the bus

part of ocean surfboard

arch over window column

the short hair of the player

the back wheel of the bus

a man wearing sunglasses

leash on dog pulling person on skateboard

racket in tennis player's hand

child is wearing pink gloves

flag on top of building

cement circle in the glass

a knife with a brown handle

silver stereo on table

umbrella covering the people

the man is surfing

a black tee shirt

a airplane is in the sky

a red and white ball

the clock says 11/08

drink can sitting on the sink

video game on the tv screen

cell phone in hand

side windows of a plane

a woman waiting at the train station

the building has windows

black screen on television

light brown cow on ground

a brown roof of a building

girl about to throw frisbee

dark shirt with yellow writing

the bear has a black nose

the clock is black

a red stripe on the plane

the red jacket of the skier

the apple is green

the bird stands in the sun

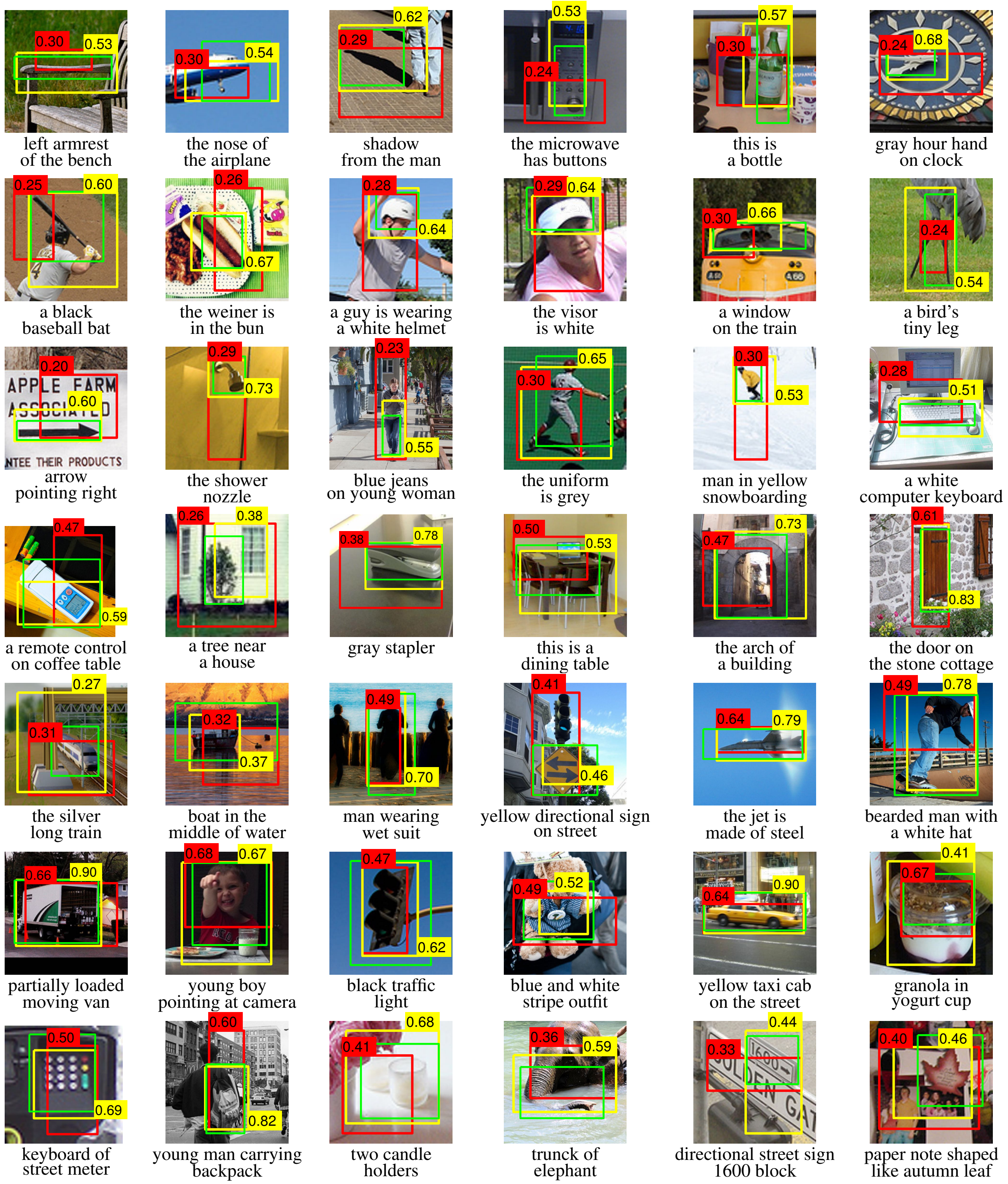
white keys of a keyboard

a letter p written in white

a fat seagull standing

Figure 6: Qualitative comparison between DBNet and DenseCap on localization task. Examples are randomly sampled. Green boxes: ground truth; Red boxes: DenseCap; Yellow boxes: DBNet. The numbers are IoU with ground truth boxes.

F.2. More qualitative comparison with SCRC



left armrest of the bench

the nose of the airplane

shadow from the man

the microwave has buttons

this is a bottle

gray hour hand on clock

a black baseball bat

the weiner is in the bun

a guy is wearing a white helmet

the visor is white

a window on the train

a bird's tiny leg

arrow pointing right

the shower nozzle

blue jeans on young woman

the uniform is grey

man in yellow snowboarding

a white computer keyboard

a remote control on coffee table

a tree near a house

gray stapler

this is a dining table

the arch of a building

the door on the stone cottage

the silver long train

boat in the middle of water

man wearing wet suit

yellow directional sign on street

the jet is made of steel

bearded man with a white hat

partially loaded moving van

young boy pointing at camera

black traffic light

blue and white stripe outfit

yellow taxi cab on the street

granola in yogurt cup

keyboard of street meter

young man carrying backpack

two candle holders

trunck of elephant

directional street sign 1600 block

paper note shaped like autumn leaf

Figure 7: Qualitative comparison between DBNet and SCRC on localization task. Examples are randomly sampled. Green boxes: ground truth; Red boxes: SCRC; Yellow boxes: DBNet. The numbers are IoU with ground truth boxes.

G. Qualitative Comparison for Detection

In this section, we showed more qualitative results for visual entity detection with various phrases. As opposed to the localization task, a decision threshold was needed to decide if the visual entity of interest exists or not. We determined this threshold either using prior knowledge on the ground truth regions (Section G.1) or based on the precision of the detector (Section G.2 and Section G.3).

In Section G.1, we showed the same number of detected regions as the ground truth regions for all methods. We visualized randomly chosen testing images and phrases under the constraint that at least one of DBNet, DenseCap, or SCRC could get sufficiently accurate detection results (IoU with a ground truh is greater than 0.4).

In Section G.2, we found a decision threshold for each text phrase to make the detection precision (for the IoU threshold at 0.5) equal to 0.5. If not applicable, we excluded that phrase from visualization. We randomly chose testing images and phrases to visualize.

In Section G.3, we used the same decision threshold as in Section G.2. However, we focused on visualizing failed detection cases. In particular, we randomly chose testing images and phrases under the constraint that at least one of DBNet, DenseCap, and SCRC gave significantly wrong detection results (IoU with any ground truth is less than 0.2). The failure types were also displayed in the figures.

See results on the next page.

G.1. Random detection results with known number of ground truths

In Figure 8, the number of ground truth entities on the image was supposed to be known in advance. All three methods (DBNet, DenseCap, and SCRC) could perform similarly for detecting queried visual entities under a loose standard for localization accuracy (e.g., counting a detected box as a true positive even if it overlaps slightly with the ground truth box). The localization accuracy of DBNet was usually more accurate.

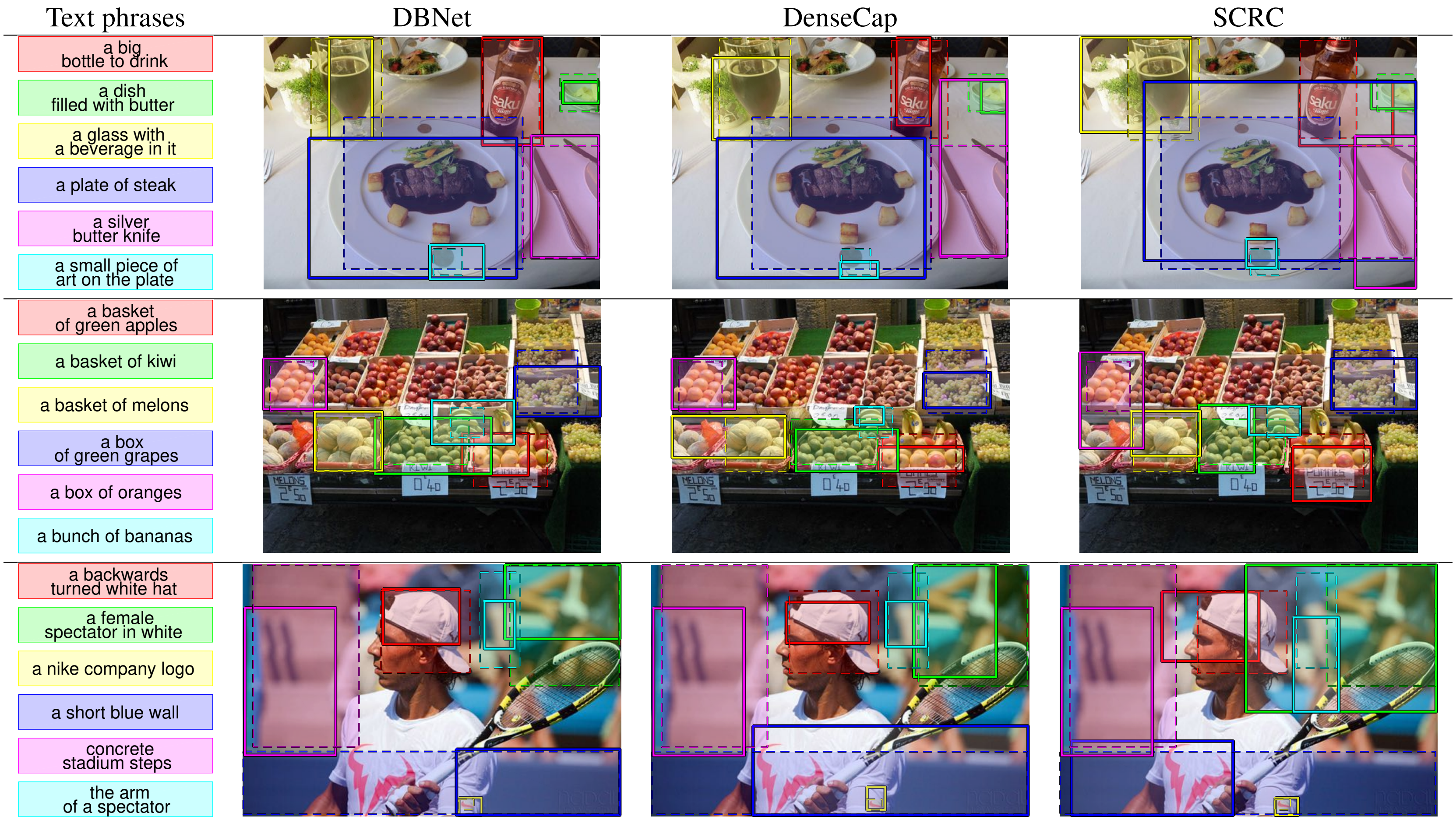


Figure 8: Qualitative detection results of DBNet, DenseCap, and SCRC when the number of ground truth is known. Detection results of six different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

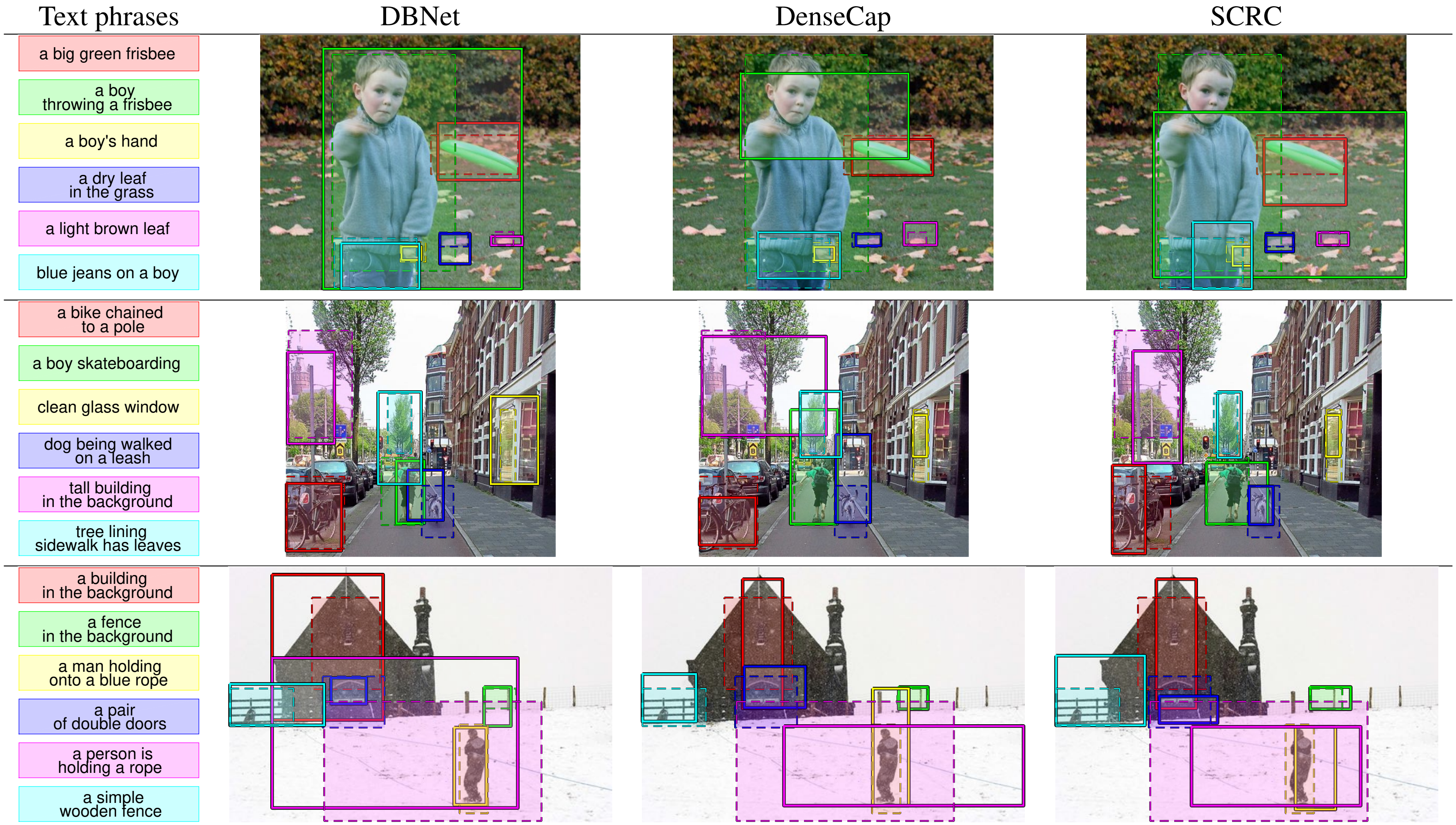


Figure 9: (continued from Figure 8) Qualitative detection results of DBNet, DenseCap, and SCRC when the number of ground truth is known. Detection results of six different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

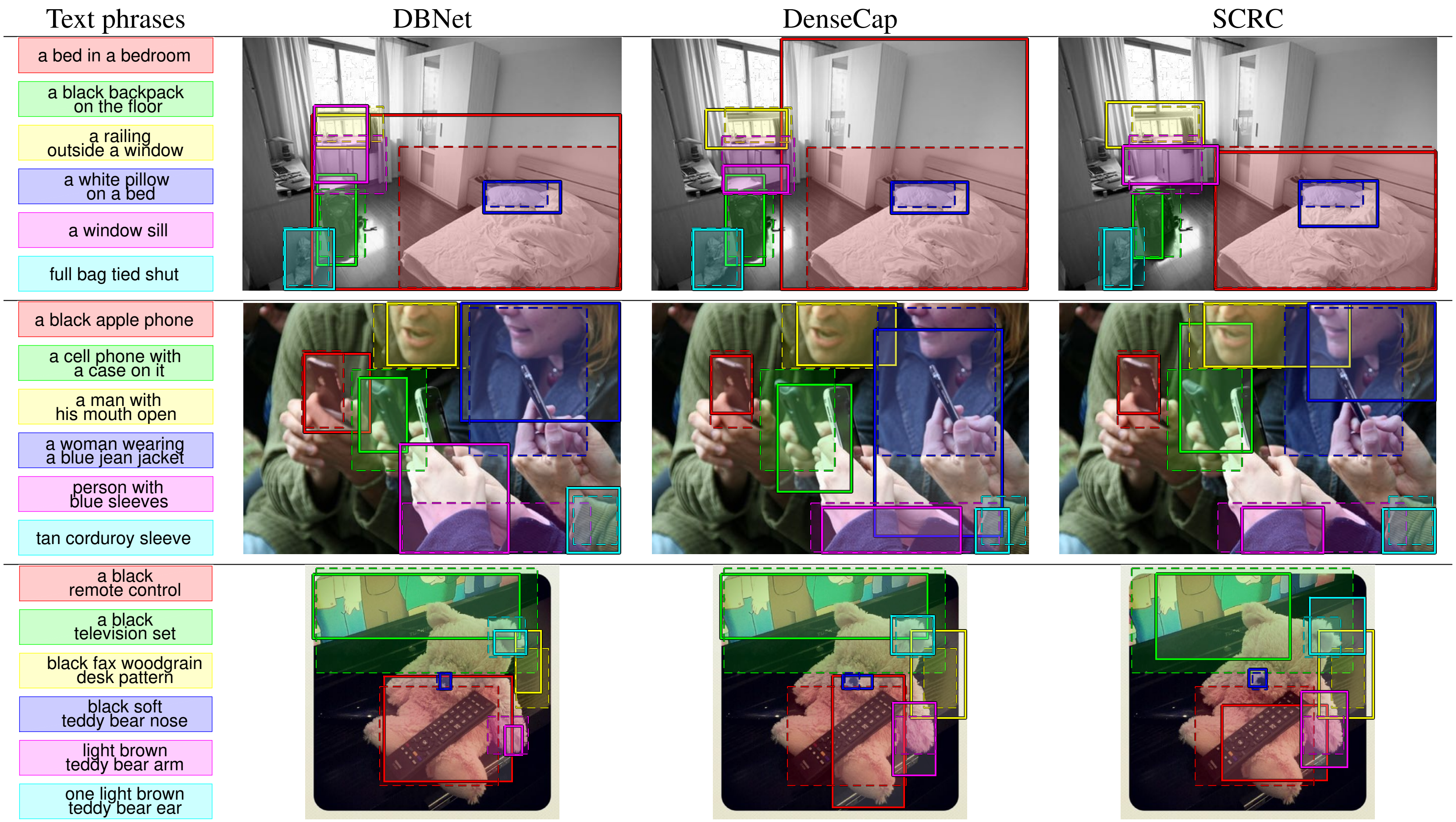

Text phrases
DBNet
DenseCap
SCRC
a bed in a bedroom
a black backpack on the floor
a railing outside a window
a white pillow on a bed
a window sill
full bag tied shut
a black apple phone
a cell phone with a case on it
a man with his mouth open
a woman wearing a blue jean jacket
person with blue sleeves
tan corduroy sleeve
a black remote control
a black television set
black fax woodgrain desk pattern
black soft teddy bear nose
light brown teddy bear arm
one light brown teddy bear ear

Figure 10: (continued from Figure 9) Qualitative detection results of DBNet, DenseCap, and SCRC when the number of ground truth is known. Detection results of six different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

G.2. Random detection results with phrase-dependent thresholds

In Figure 11, we used phrase-dependent decision thresholds to determine how many regions were detected on an image. We set the threshold to make the detection precision for the IoU threshold at 0.5 equal to 0.5 when applicable. DBNet outperformed DenseCap and SCRC significantly. DenseCap and SCRC resulted in many cases of false alarms or miss detection. Note that DBNet could usually achieve the 0.5 precision with a reasonable recall level, but DenseCap and SCRC might either fail achieving the 0.5 precision at all or give a low recall.

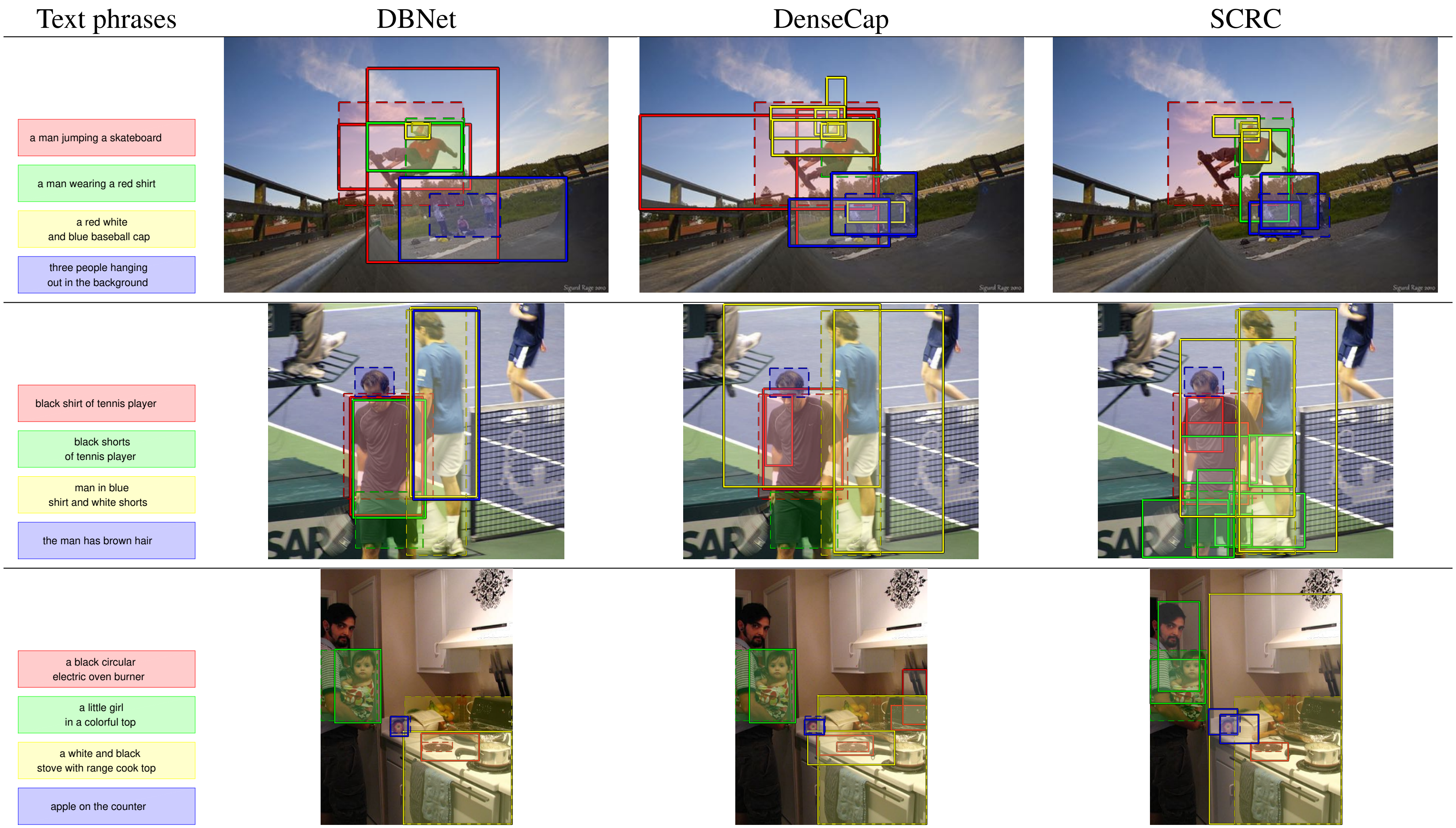

Text phrases
DBNet
DenseCap
SCRC
a man jumping a skateboard
a man wearing a red shirt
a red white and blue baseball cap
three people hanging out in the background
black shirt of tennis player
black shorts of tennis player
man in blue shirt and white shorts
the man has brown hair
a black circular electric oven burner
a little girl in a colorful top
a white and black stove with range cook top
apple on the counter

Figure 11: Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

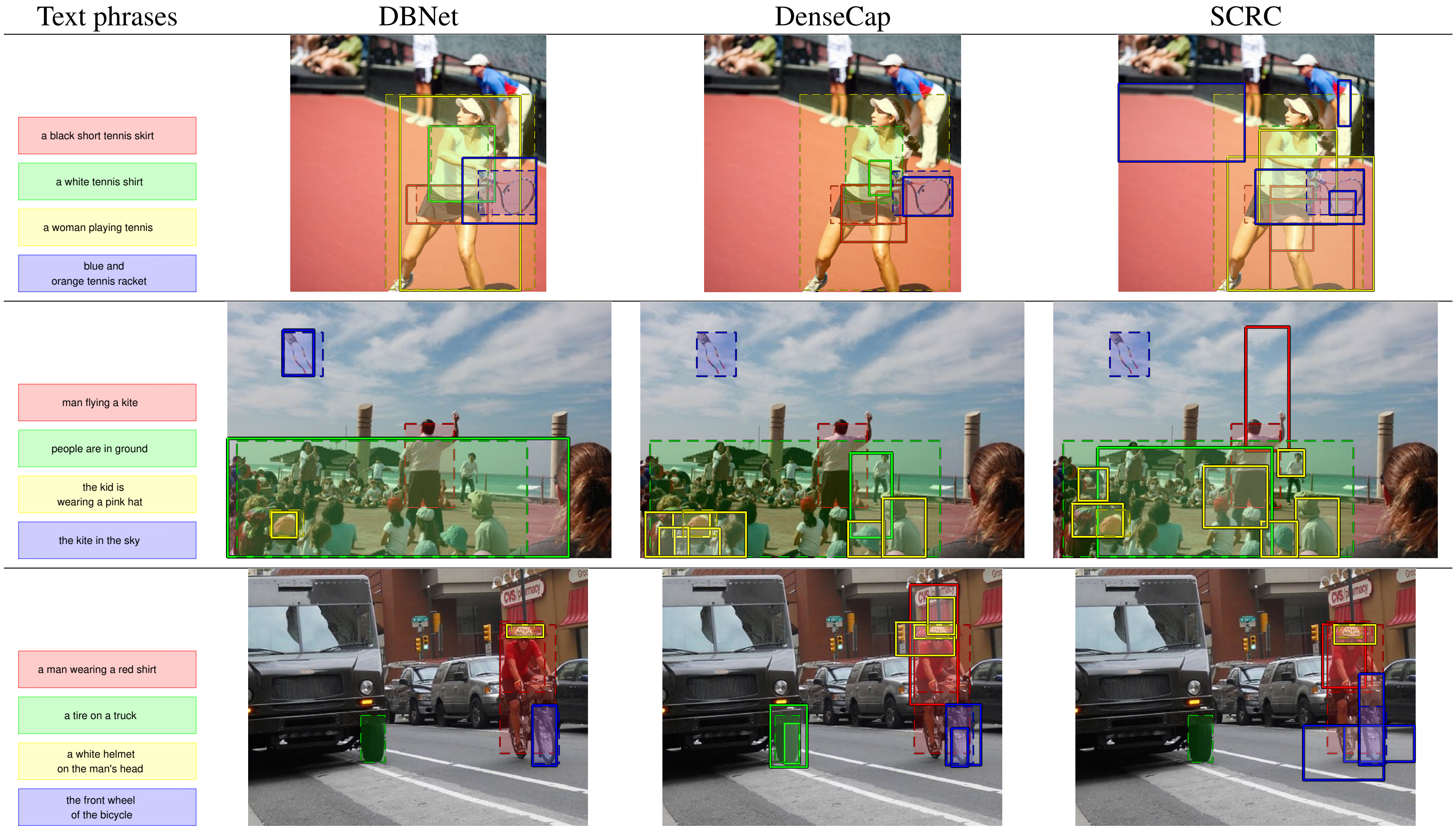

Text phrases
DBNet
DenseCap
SCRC
a black short tennis skirt
a white tennis shirt
a woman playing tennis
blue and orange tennis racket
man flying a kite
people are in ground
the kid is wearing a pink hat
the kite in the sky
a man wearing a red shirt
a tire on a truck
a white helmet on the man's head
the front wheel of the bicycle

Figure 12: (continued from Figure 11) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

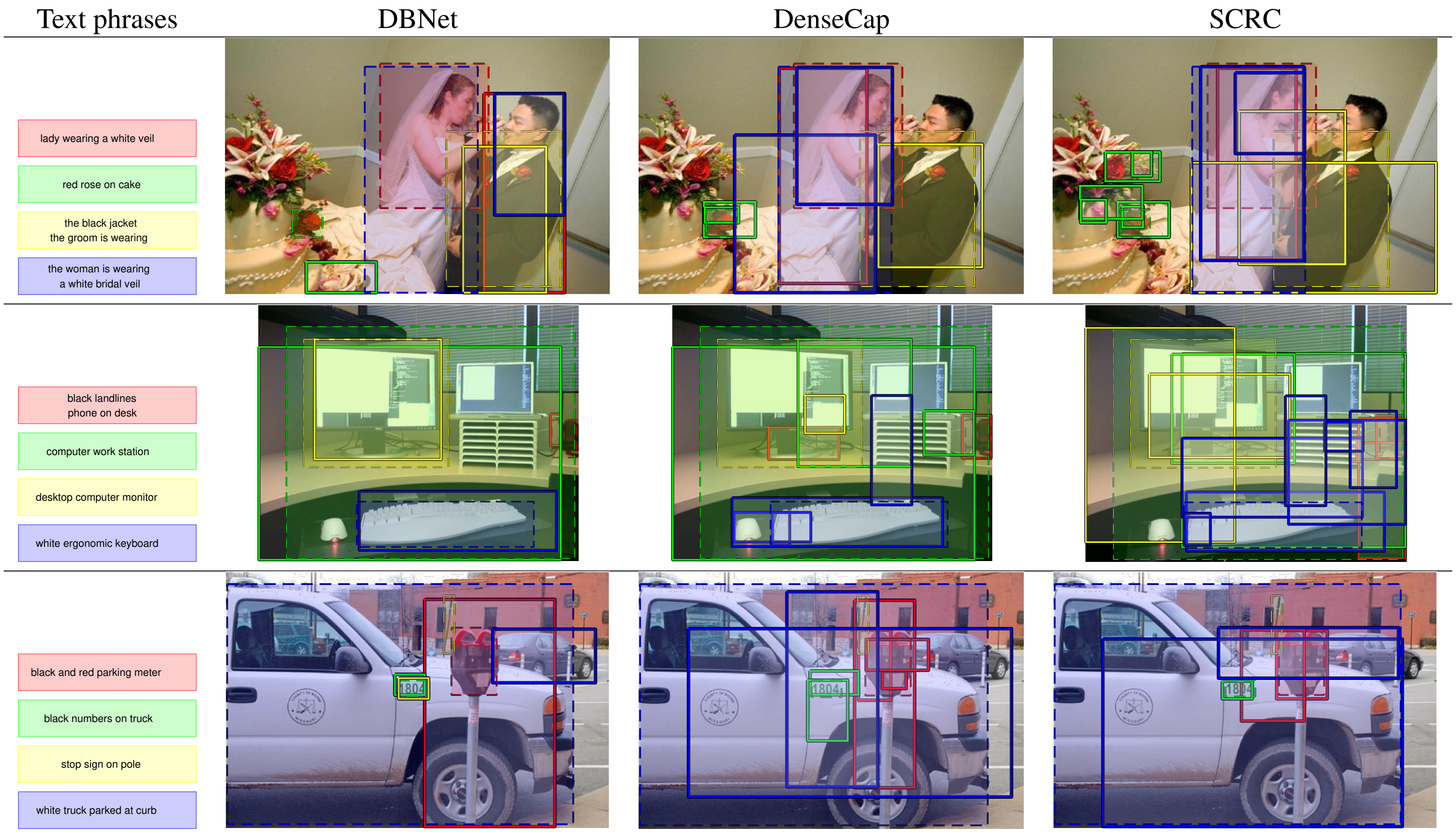

Text phrases
DBNet
DenseCap
SCRC
lady wearing a white veil
red rose on cake
the black jacket the groom is wearing
the woman is wearing a white bridal veil
black landlines phone on desk
computer work station
desktop computer monitor
white ergonomic keyboard
black and red parking meter
black numbers on truck
stop sign on pole
white truck parked at curb
1804

Figure 13: (continued from Figure 12) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

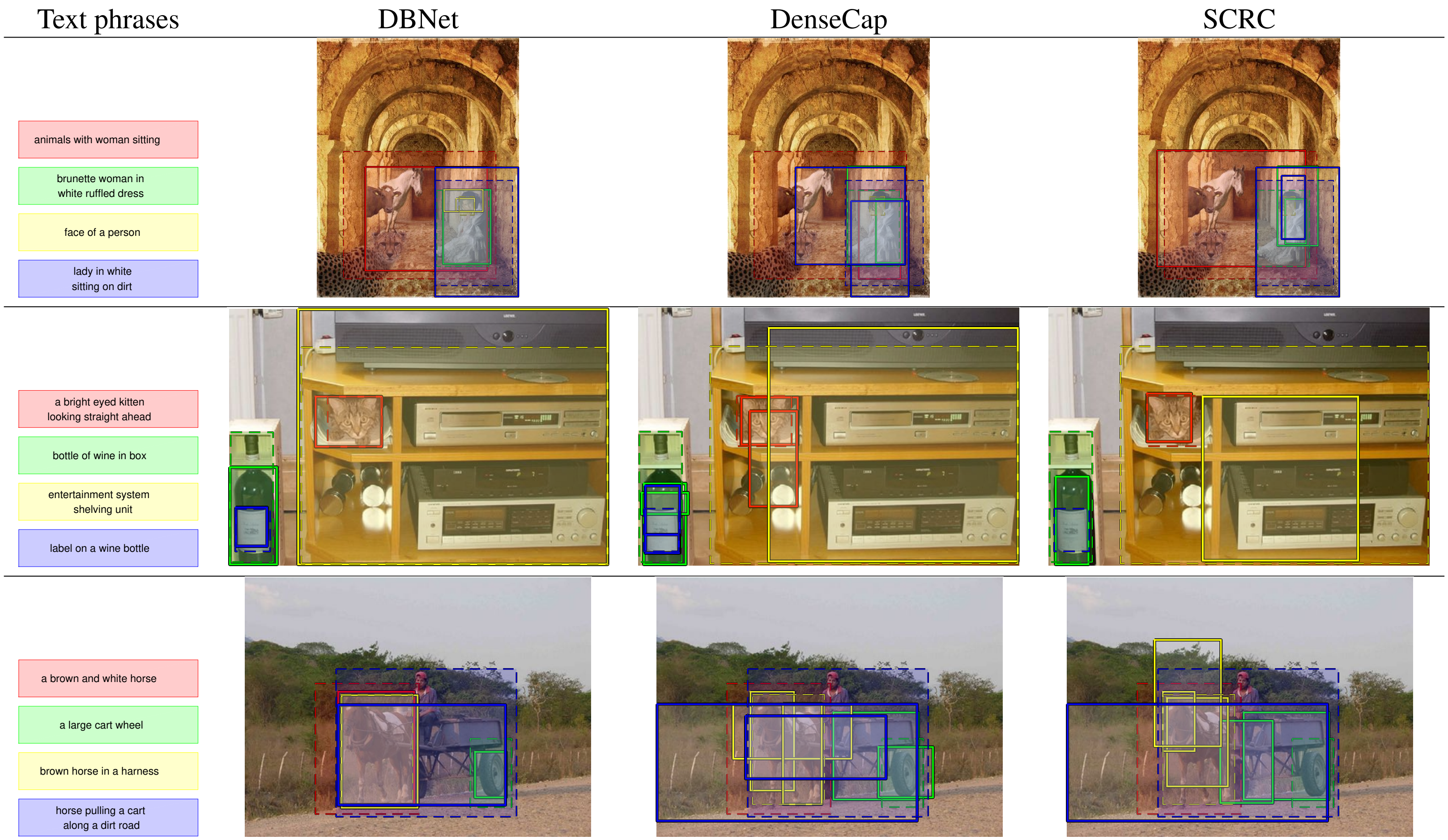

Text phrases
DBNet
DenseCap
SCRC
animals with woman sitting
brunette woman in white ruffled dress
face of a person
lady in white sitting on dirt
a bright eyed kitten looking straight ahead
bottle of wine in box
entertainment system shelving unit
label on a wine bottle
a brown and white horse
a large cart wheel
brown horse in a harness
horse pulling a cart along a dirt road

Figure 14: (continued from Figure 13) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

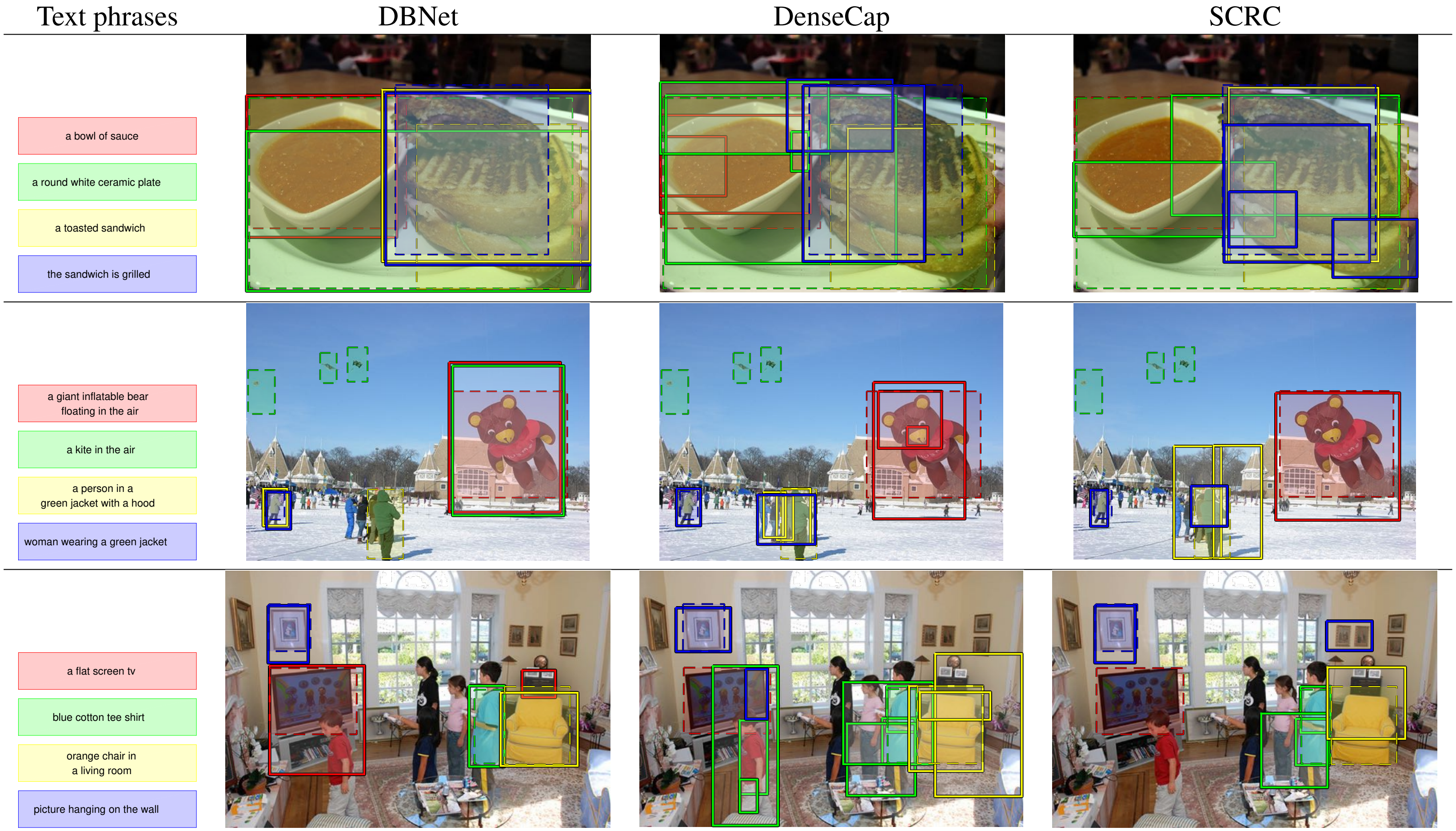

Text phrases
DBNet
DenseCap
SCRC
a bowl of sauce
a round white ceramic plate
a toasted sandwich
the sandwich is grilled
a giant inflatable bear floating in the air
a kite in the air
a person in a green jacket with a hood
woman wearing a green jacket
a flat screen tv
blue cotton tee shirt
orange chair in a living room
picture hanging on the wall

Figure 15: (continued from Figure 14) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

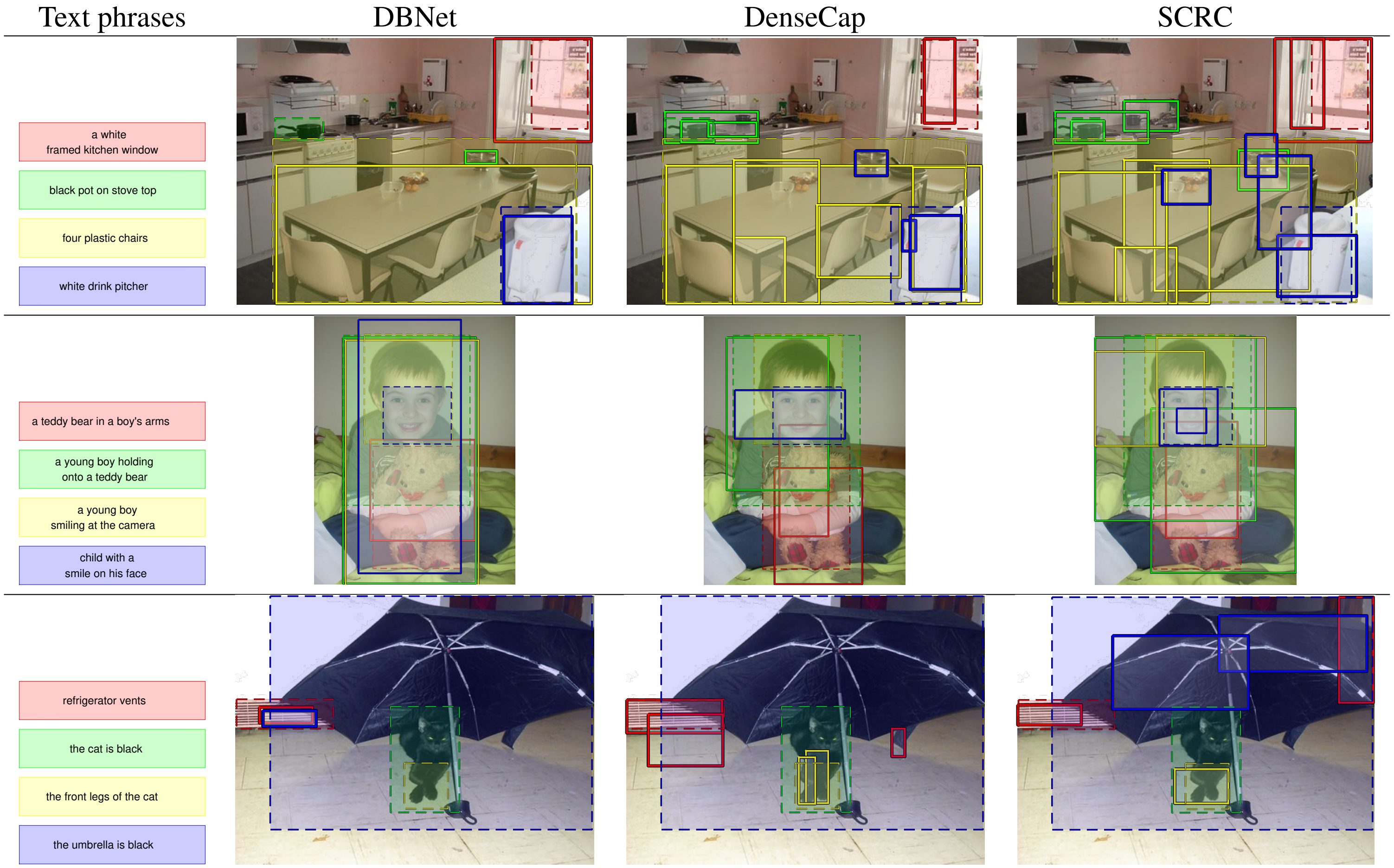

Text phrases
DBNet
DenseCap
SCRC
a white framed kitchen window
black pot on stove top
four plastic chairs
white drink pitcher
a teddy bear in a boy's arms
a young boy holding onto a teddy bear
a young boy smiling at the camera
child with a smile on his face
refrigerator vents
the cat is black
the front legs of the cat
the umbrella is black

Figure 16: (continued from Figure 15) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

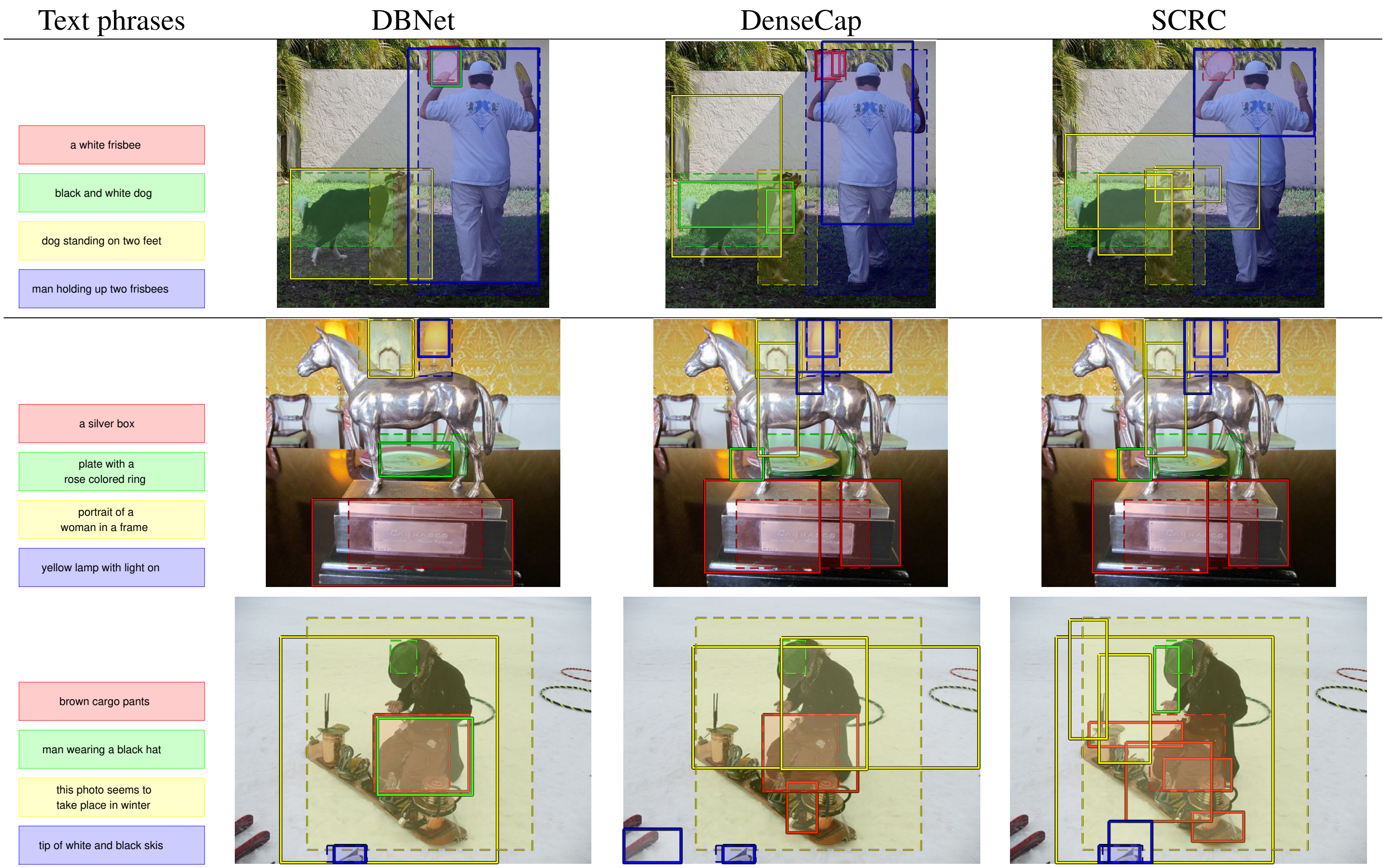

Text phrases
DBNet
DenseCap
SCRC
a white frisbee
black and white dog
dog standing on two feet
man holding up two frisbees
a silver box
plate with a rose colored ring
portrait of a woman in a frame
yellow lamp with light on
brown cargo pants
man wearing a black hat
this photo seems to take place in winter
tip of white and black skis

Figure 17: (continued from Figure 16) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

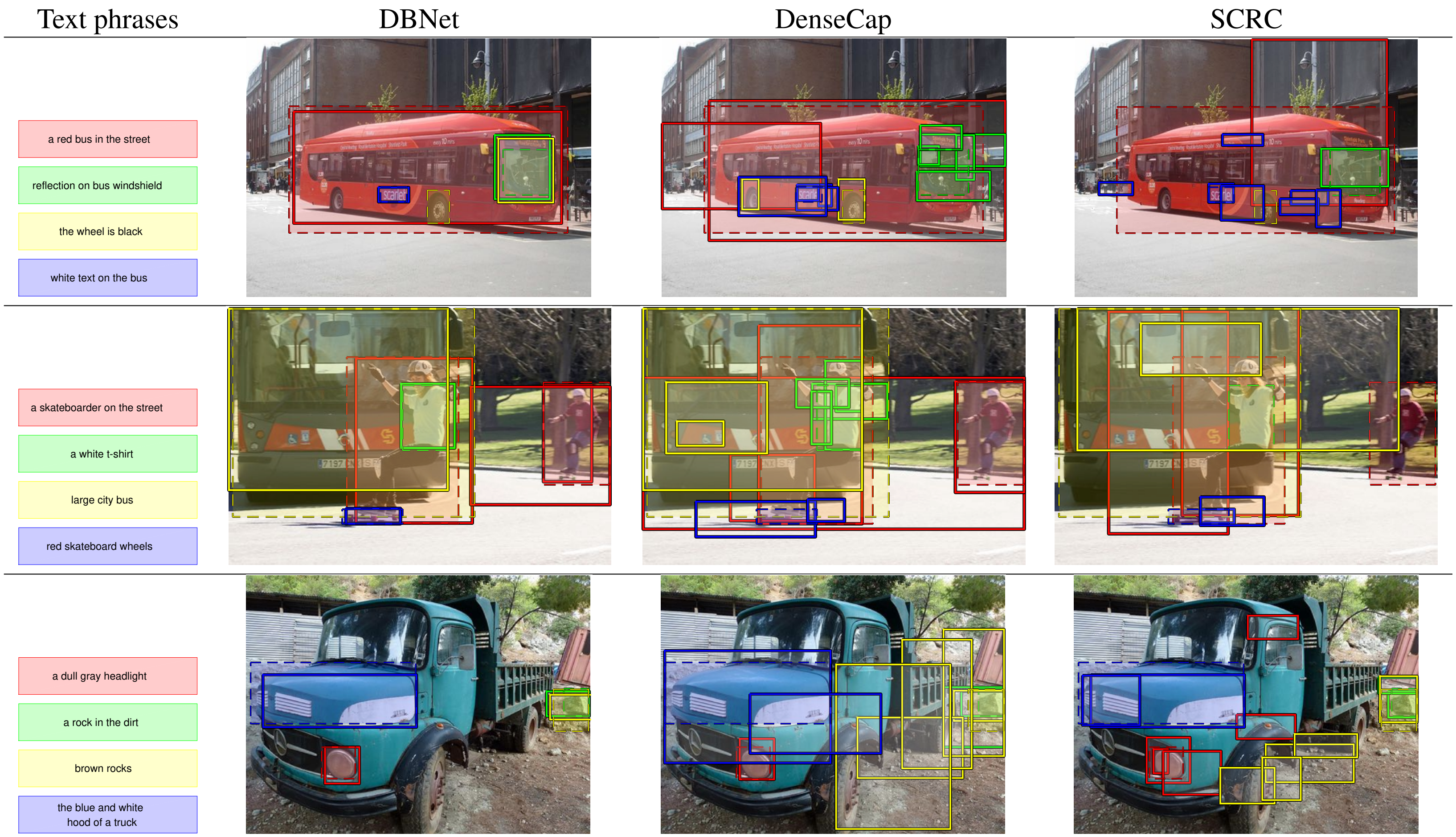


Figure 18: (continued from Figure 17) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

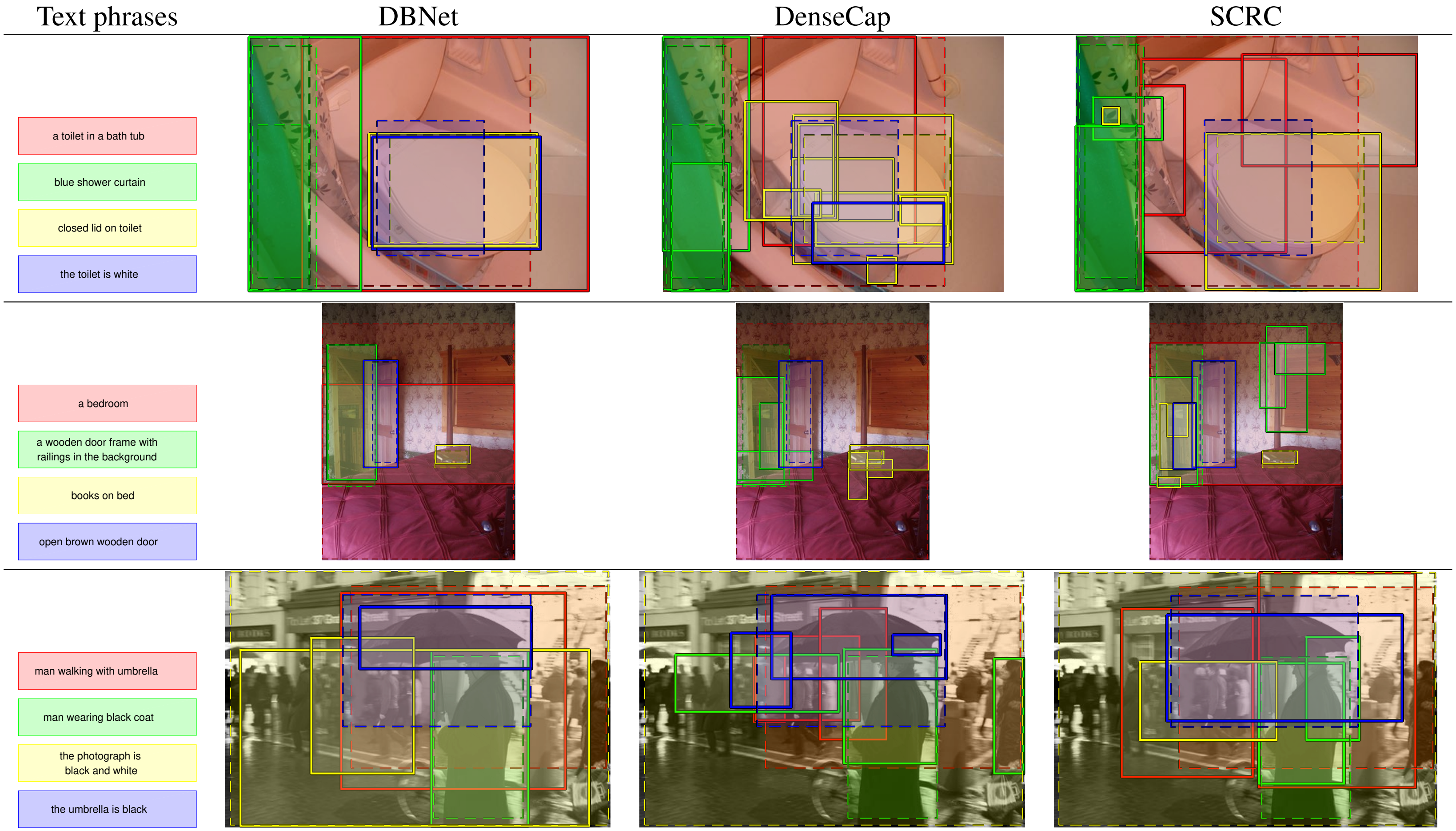


Figure 19: (continued from Figure 18) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

G.3. Failure cases for detection with phrase-dependent thresholds

In this section, we used phrase-dependent decision thresholds in the same way as in Section G.2, except for focusing on showing failure cases. We visualized randomly chosen testing images and phrases under the constraint that at least one of DBNet, DenseCap, and SCRC should significantly fail in detection (i.e., IoU with ground truth is less than 0.2). In Figure 20, we categorized failure cases into three types: 1) the false alarm (the detected box has no overlap with any ground truth), 2) inaccurate localization (the IoU with ground truth is less than 0.5), 3) missing detection (no detection box has overlap with a ground truth region). For each image, we showed only one phrase for visual clarity and displayed the failure types for comprehensiveness. DBNet has significantly less failure cases than DenseCap and SCRC.

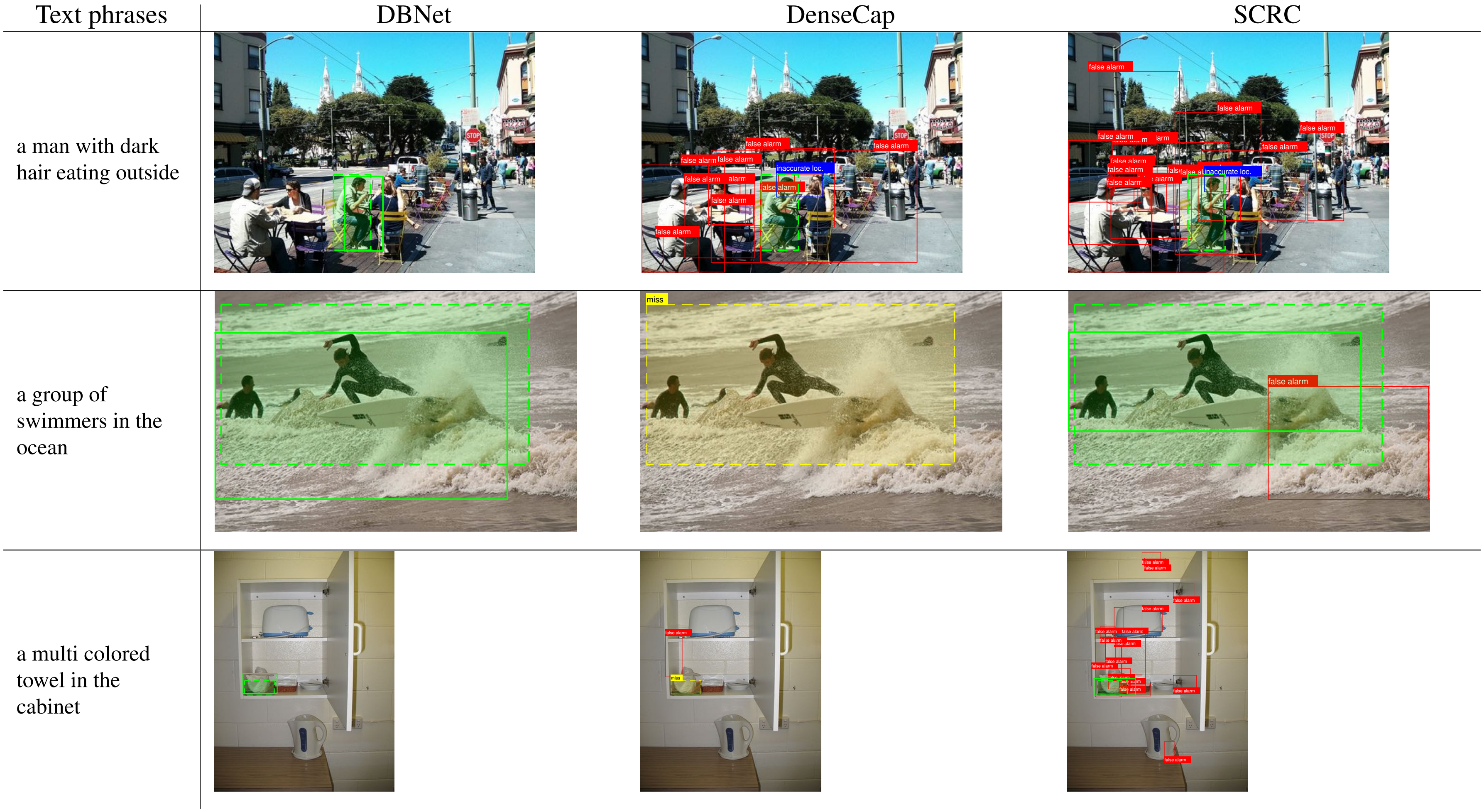

Text phrases
DBNet
DenseCap
SCRC
a man with dark hair eating outside
false alarm
inaccurate loc.
a group of swimmers in the ocean
miss
false alarm
a multi colored towel in the cabinet
false alarm
miss

Figure 20: Random failure examples. Green boxes with solid boundary: successful detection ($IoU \geq 0.5$); Green boxes with dashed boundary: ground truth with matched detection; Red boxes: false alarm; Yellow boxes with dashed boundary: missed ground truth (without matched detection); Blue boxes: inaccurately localized detection ($0 < IoU < 0.5$).

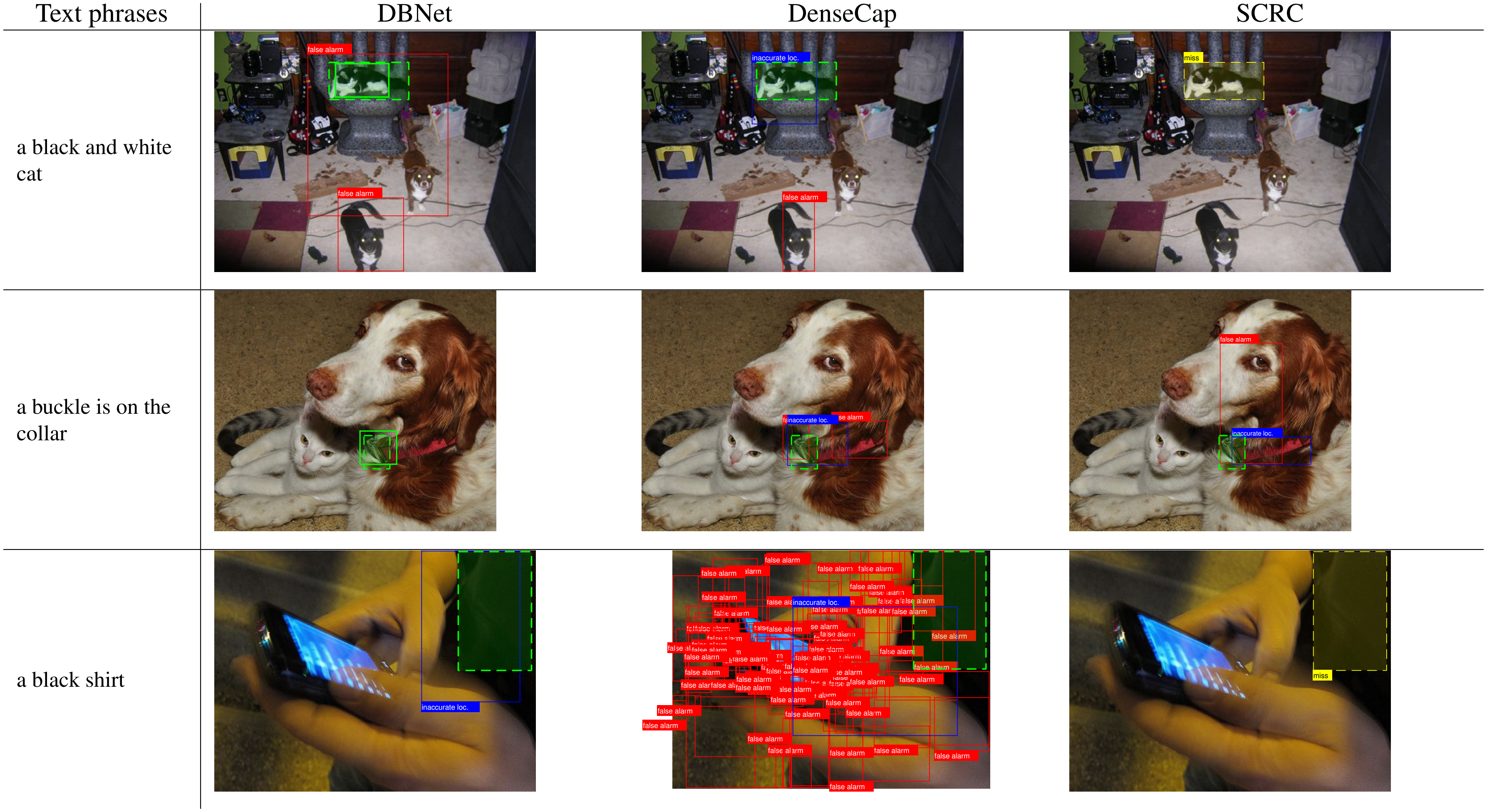


Figure 21: (continued from Figure 20) Random failure examples. Green boxes with solid boundary: successful detection ($IoU \geq 0.5$); Green boxes with dashed boundary: ground truth with matched detection; Red boxes: false alarm; Yellow boxes with dashed boundary: missed ground truth (without matched detection); Blue boxes: inaccurately localized detection ($0 < IoU < 0.5$).

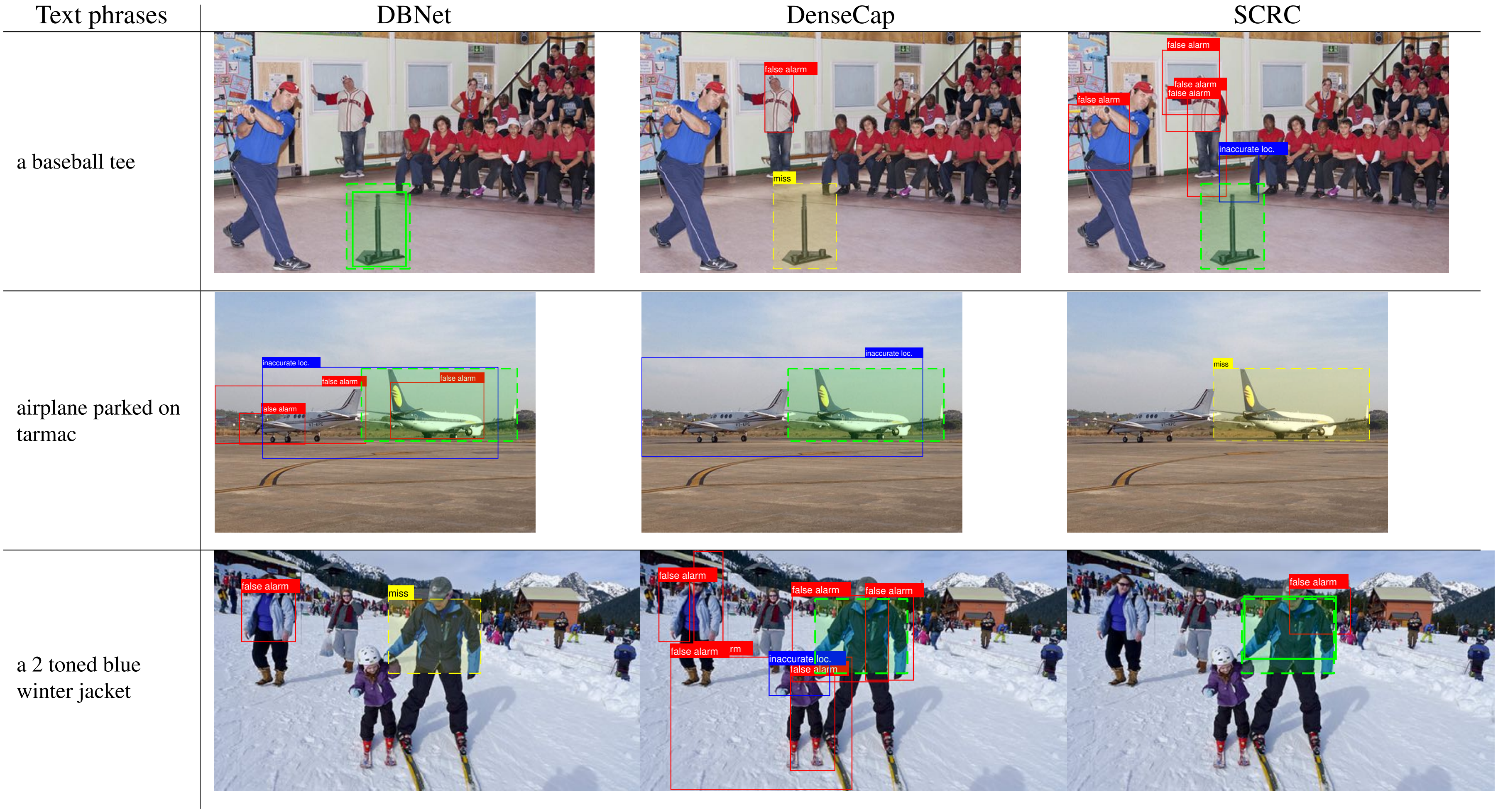


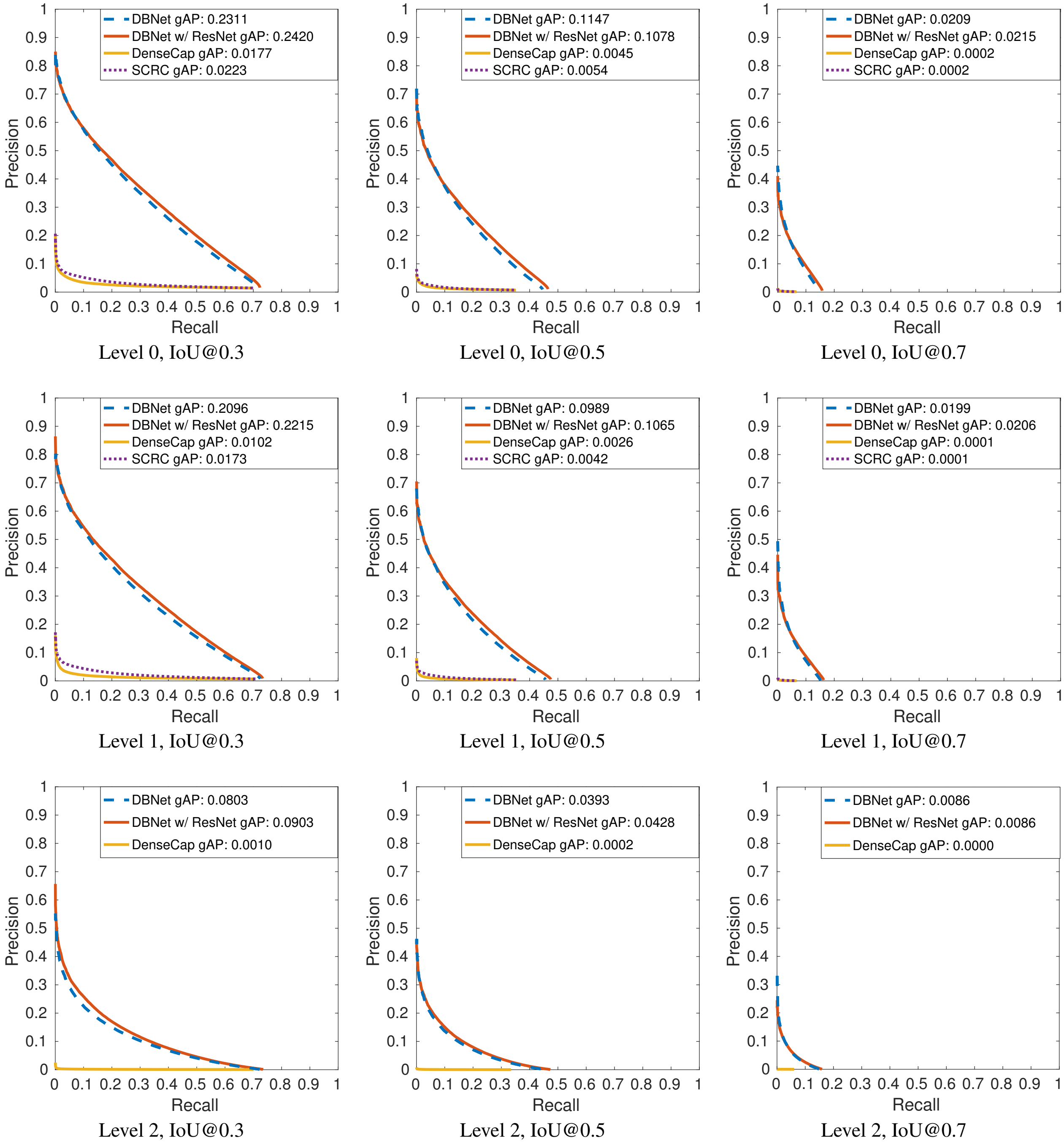
Figure 22: (continued from Figure 21) Random failure examples. Green boxes with solid boundary: successful detection ($IoU \geq 0.5$); Green boxes with dashed boundary: ground truth with matched detection; Red boxes: false alarm; Yellow boxes with dashed boundary: missed ground truth (without matched detection); Blue boxes: inaccurately localized detection ($0 < IoU < 0.5$).

H. Precision-recall curves

We show precision-recall curves for both global average precision (gAP) (Section H.1) and mean average precision (mAP) (Section H.2) calculation.

H.1. Phrase-independent precision-recall curves

We reported precision-recall curves for different query set under different IoU threshold using the detection results for all test cases in Figure 23. gAP was computed based on these precision-recall curves.



Level 0, IoU@0.3

Level 0, IoU@0.5

Level 0, IoU@0.7

Level 1, IoU@0.3

Level 1, IoU@0.5

Level 1, IoU@0.7

Level 2, IoU@0.3

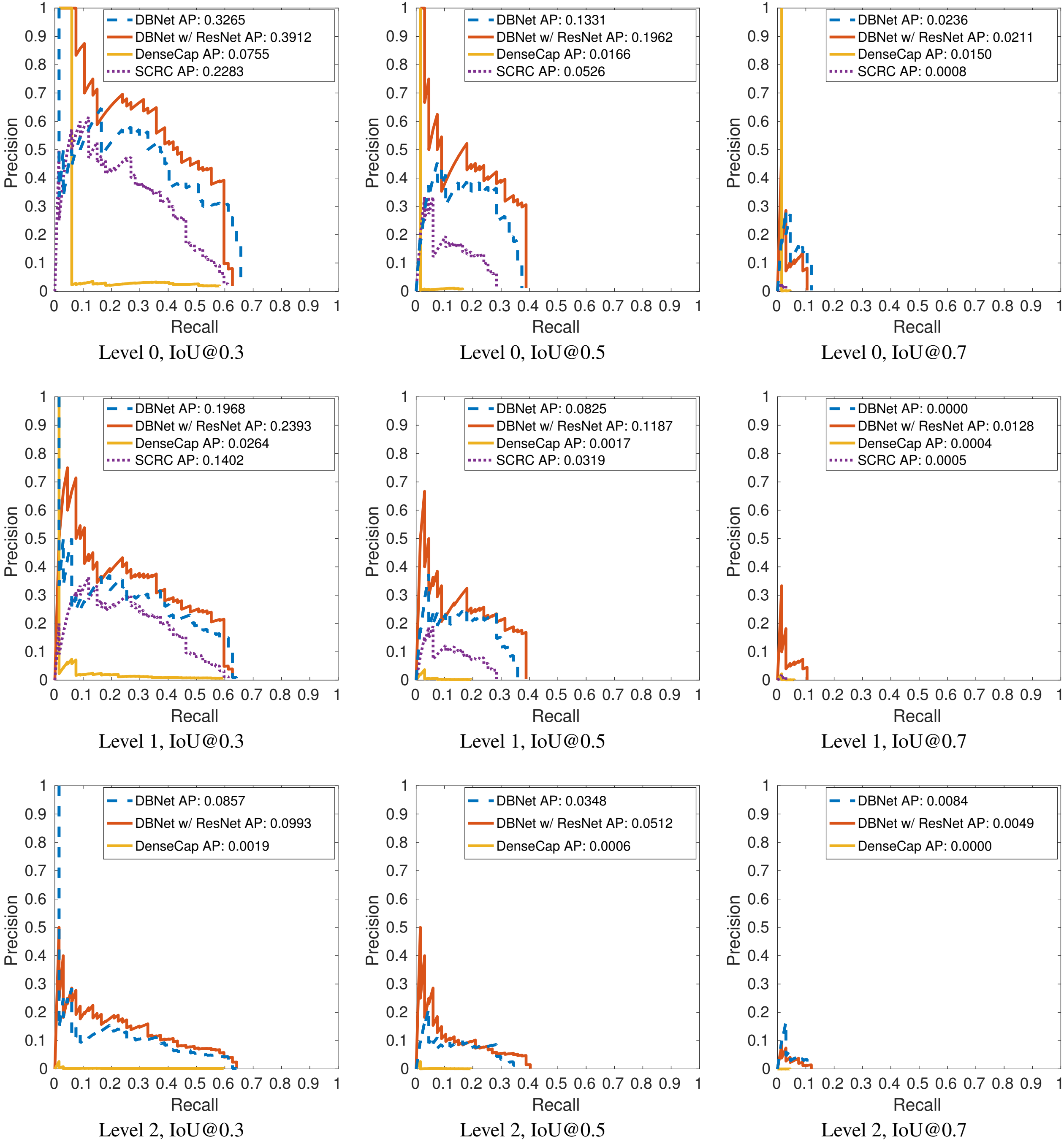
Level 2, IoU@0.5

Level 2, IoU@0.7

Figure 23: Phrase-independent precision-recall curves for calculating gAP.

H.2. Phrase-dependent precision-recall curves

We calculated precision-recall curves using various query sets under different IoU thresholds independently for different text phrases over the entire test set. mAP was computed based on these precision-recall curves. We showed precision-recall curves for a few selected text phrases in Figure 24, 25, 26, 27, and 28.



Level 0, IoU@0.3

Level 0, IoU@0.5

Level 0, IoU@0.7

Level 1, IoU@0.3

Level 1, IoU@0.5

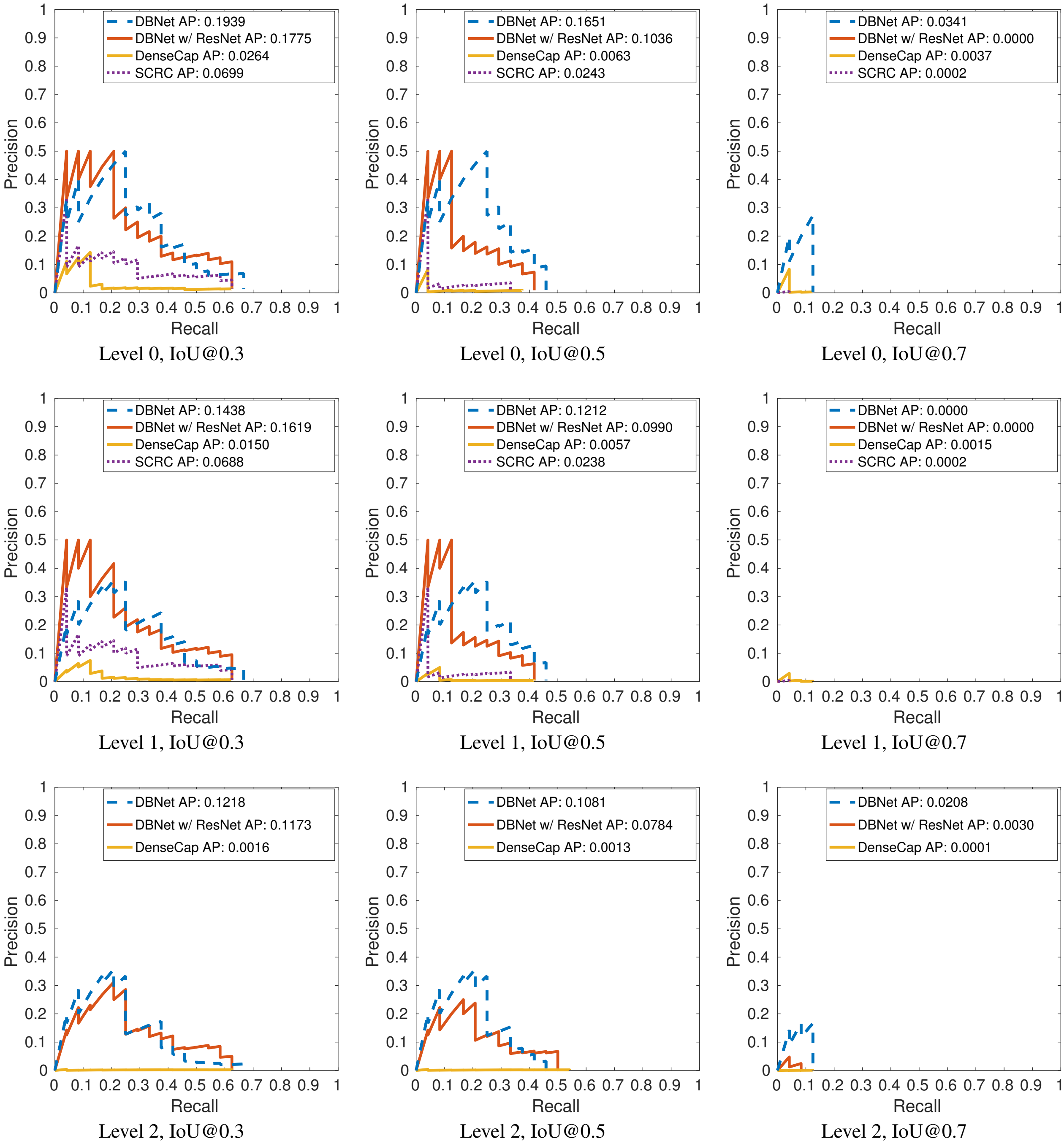
Level 1, IoU@0.7

Level 2, IoU@0.3

Level 2, IoU@0.5

Level 2, IoU@0.7

Figure 24: Precision-recall curves for text phrase “head of a person”.



Level 0, IoU@0.3

Level 0, IoU@0.5

Level 0, IoU@0.7

Level 1, IoU@0.3

Level 1, IoU@0.5

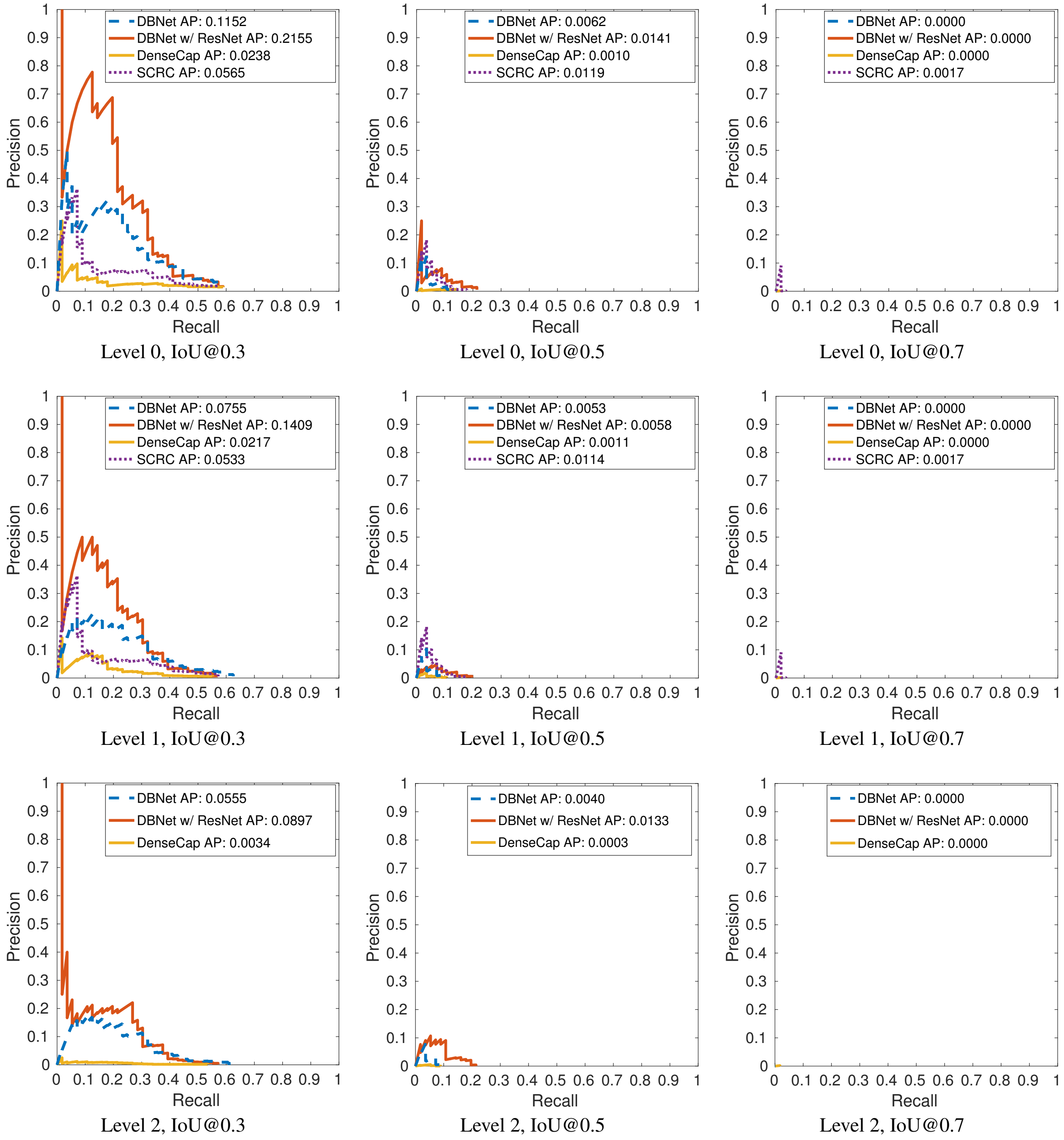
Level 1, IoU@0.7

Level 2, IoU@0.3

Level 2, IoU@0.5

Level 2, IoU@0.7

Figure 25: Precision-recall curves for text phrase “a window on the building”.



Level 0, IoU@0.3

Level 0, IoU@0.5

Level 0, IoU@0.7

Level 1, IoU@0.3

Level 1, IoU@0.5

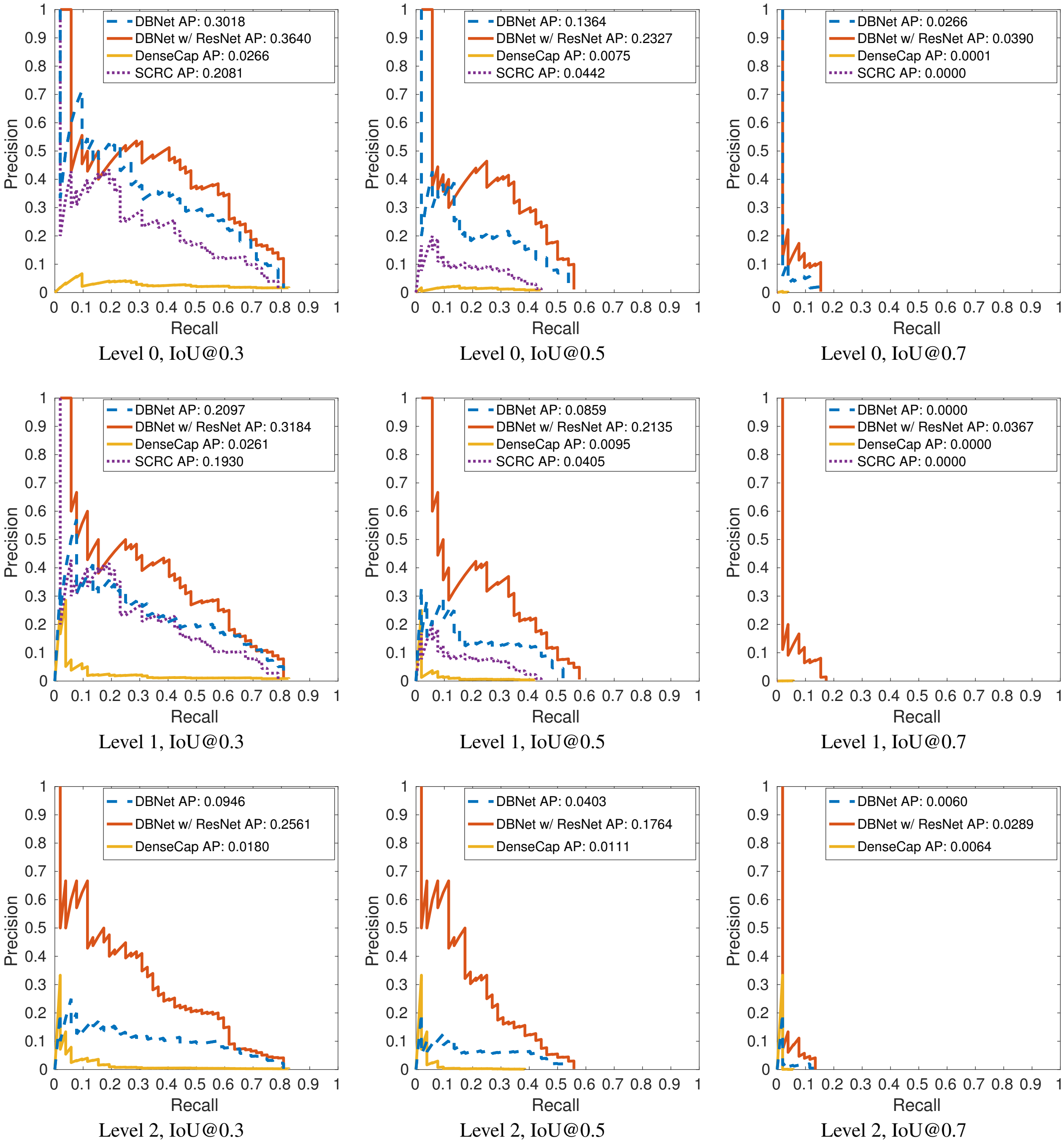
Level 1, IoU@0.7

Level 2, IoU@0.3

Level 2, IoU@0.5

Level 2, IoU@0.7

Figure 26: Precision-recall curves for text phrase “the water is calm”.



Level 0, IoU@0.3

Level 0, IoU@0.5

Level 0, IoU@0.7

Level 1, IoU@0.3

Level 1, IoU@0.5

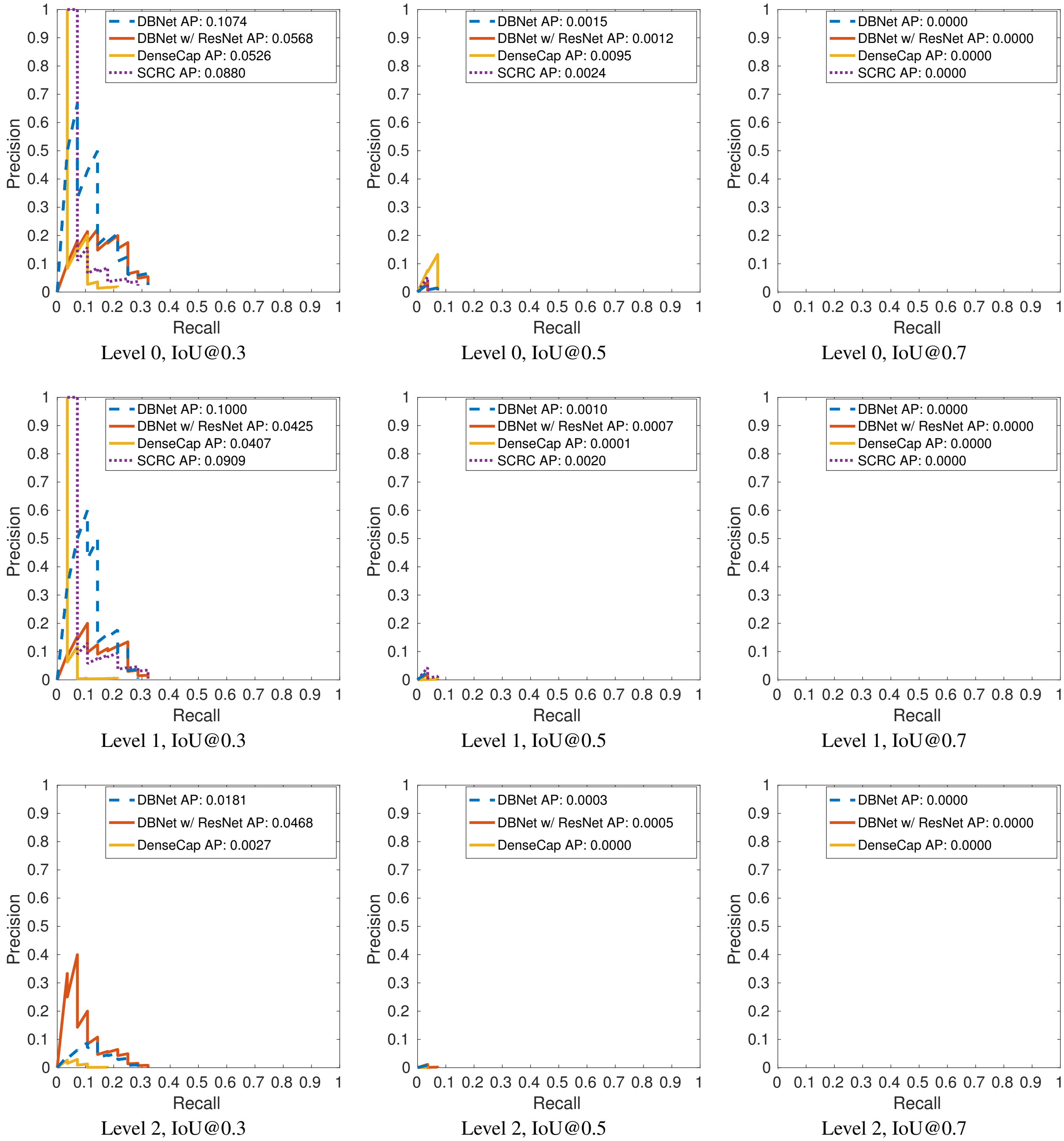
Level 1, IoU@0.7

Level 2, IoU@0.3

Level 2, IoU@0.5

Level 2, IoU@0.7

Figure 27: Precision-recall curves for text phrase “man wearing blue jeans”.



Level 0, IoU@0.3

Level 0, IoU@0.5

Level 0, IoU@0.7

Level 1, IoU@0.3

Level 1, IoU@0.5

Level 1, IoU@0.7

Level 2, IoU@0.3

Level 2, IoU@0.5

Level 2, IoU@0.7

Figure 28: Precision-recall curves for text phrase “small ripples in the water”.