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

Abstract

*Associating image regions with text queries has been recently explored as a new way to bridge visual and linguistic representations. A few pioneering approaches have been proposed based on recurrent neural language models trained generatively (e.g., generating captions), but achieving somewhat limited localization accuracy. To better address natural-language-based visual entity localization, we propose a discriminative approach. We formulate a discriminative bimodal neural network (**DBNet**), which can be trained by a classifier with extensive use of negative samples. Our training objective encourages better localization on single images, incorporates text phrases in a broad range, and properly pairs image regions with text phrases into positive and negative examples. Experiments on the Visual Genome dataset demonstrate the proposed DBNet significantly outperforms previous state-of-the-art methods both for localization on single images and for detection on multiple images. We we also establish an evaluation protocol for natural-language visual detection.*

1. Introduction

Object localization and detection in computer vision are traditionally limited to a small number of predefined categories (e.g., car, dog, and person), and category-specific image region classifiers [7, 11, 14] serve as object detectors. However, in the real world, the *visual entities* of interest are much more diverse, including groups of objects (involved in certain relationships), object parts, and objects with particular attributes and/or in particular context. For scalable annotation, these entities need to be labeled in a more flexible way, such as using text phrases.

Deep learning has been demonstrated as a unified learning framework for both text and image representations. Significant progress has been made in many related tasks, such as image captioning [55, 56, 25, 37, 5, 9, 23, 18, 38], visual question answering [3, 36, 57, 41, 2], text-based fine-
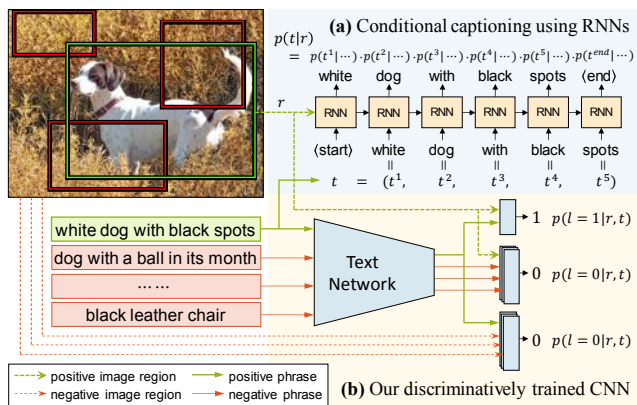

Figure 1: Comparison between (a) image captioning model and (b) our discriminative architecture for visual localization.

grained image classification [44], natural-language object retrieval [21, 38], and text-to-image generation [45].

A few pioneering works [21, 38] use recurrent neural language models [15, 39, 50] and deep image representations [31, 49] for localizing the object referred to by a text phrase given a single image (i.e., "object referring" task [26]). Global spatial context, such as "a man on the left (of the image)", has been commonly used to pick up the particular object. In contrast, Johnson et al. [23] takes descriptions without global context[1] as queries for localizing more general visual entities on the Visual Genome dataset [30].

All above existing work performs localization by maximizing the likelihood to generate the query text given image regions using an image captioning model (Figure 1a), whose output probability density needs to be modeled on the virtually infinite space of the natural language. Since it is hard to train a classifier on such a huge structured output space, current captioning models are constrained to be trained in generative [21, 23] or partially discriminative [38] ways. However, as discriminative tasks, localization and detection usually favor models that are trained with a more

---

[1]Only a very small portion of text phrases on the Visual Genome refer to the global context.

discriminative objective to better utilize negative samples.

In this paper, we propose a new deep architecture for natural-language-based visual entity localization, which we call a *discriminative bimodal network* (**DBNet**). Our architecture uses a binary output space to allow extensive discriminative training, where any negative training sample can be potentially utilized. The key idea is to take the text query as a condition rather than an output and to let the model directly predict if the text query and image region are compatible (Figure 1b). In particular, the two pathways of the deep architecture respectively extract the visual and linguistic representations. A discriminative pathway is built upon the two pathways to fuse the bimodal representations for binary classification of the inter-modality compatibility.

Compared to the estimated probability density in the huge space of the natural language, the score given by a binary classifier is more likely to be *calibrated*. In particular, better calibrated scores should be more comparable across different images and text queries. This property makes it possible to learn decision thresholds to determine the existence of visual entities on multiple images and text queries, making the localization model generalizable for detection tasks. While a few examples of natural-language visual detection are showcased in [23], we perform more comprehensive quantitive and ablative evaluations.

In our proposed architecture, we use convolutional neural networks (CNNs) for both visual and textual representations. Inspired by fast R-CNN [13], we use the RoI-pooling architecture induced from large-scale image classification networks for efficient feature extraction and model learning on image regions. For textual representations, we develop a character-level CNN [60] for extracting phrase features. A network on top of the image and language pathways *dynamically* forms classifiers for image region features depending on the text features, and it outputs the classifier responses on all regions of interest.

Our main contributions are as follows:

1. We develop a bimodal deep architecture with a binary output space to enable fully discriminative training for natural-language visual localization and detection.

2. We propose a training objective that extensively pairs text phrases and bounding boxes, where 1) the discriminative objective is defined over all possible region-text pairs in the entire training set, and 2) the non-mutually exclusive nature of text phrases is taken into account to avoid ambiguous training samples.

3. Experimental results on Visual Genome demonstrate that the proposed DBNet significantly outperforms existing methods based on recurrent neural language models for visual entity localization on single images.

4. We also establish evaluation methods for natural-language visual detection on multiple images and show state-of-the-art results.

2. Related work

Object detection. Recent success of deep learning on visual object recognition [31, 59, 49, 51, 53, 17] constitutes the backbone of the state-of-the-art for object detection [14, 48, 52, 61, 42, 43, 13, 46, 17, 6]. Natural-language visual detection can adapt the deep visual representations and single forward-pass computing framework (e.g., RoI pooling [13], SPP [16], R-FCN [6]) used in existing work of traditional object detection. However, natural-language visual detection needs a huge structured label space to represent the natural language, and finding a proper mapping to the huge space from visual representations is difficult.

Image captioning and caption grounding. The recurrent neural network (RNN) [19] based language model [15, 39, 50] has become the dominant method for captioning images with text [55]. Despite differences in details of network architectures, most RNN language models learn the likelihood of picking up a word from a predefined vocabulary given the visual appearance features and previous words (Figure 1a). Xu et al. [56] introduced an attention mechanism to encourage RNNs to focus on relevant image regions when generating particular words. Karpathy and Fei-Fei [25] used strong supervision of text-region alignment for well-grounded captioning.

Object localization by natural language. Recent work used the conditional likelihood of captioning an image region with given text for localizing associated objects. Hu et al. [21] proposed the spatial-context recurrent ConvNet (SCRC), which conditioned on both local visual features and global contexts for evaluating given captions. Johnson et al. [23] combined captioning and object proposal in an end-to-end neural network, which can densely caption (DenseCap) image regions and localize objects. Mao et al. [38] trained the captioning model by maximizing the posterior of localizing an object given the text phrase, which reduced the ambiguity of generated captions. However, the training objective was limited to figuring out single objects on single images. Lu et al. [34] simplified and limited text queries to subject-relationship-object (SVO) triplets. Rohrbach et al. [47] improved localization accuracy with an extra text reconstruction task. Hu et al. [20] extended bounding box localization to instance segmentation using natural language queries. Yu et al. [58] and Nagaraja et al. [40] explicitly modeled context for referral expressions.

Text representation. Neural networks can also embed text into a fixed-dimensional feature space. Most RNN-based methods (e.g., skip-thought vectors [29]) and CNN-based methods [24, 27] use word-level one-hot encoding as the input. Recently, character-level CNN has also been demonstrated an effective way for paragraph categorization [60] and zero-shot image classification [44].

3. Discriminative visual-linguistic network

The best-performing object detection framework [7, 11, 14] in terms of accuracy generally verifies if a candidate image region belongs to a particular category of interest. Though recent deep architectures [52, 46, 23] can propose regions with confidence scores at the same time, a verification model, taking as input the image features from the exact proposed regions, still serves as a key to boost the accuracy.

In this section, we develop a verification model for natural-language visual localization and detection. Unlike the classifiers for a small number of predefined categories in traditional object detection, our model is dynamically adaptable to different text phrases.

3.1. Model framework

Let x be an image, r be the coordinates of a region, and t be a text phrase. The verification model $f(x, t, r; \Theta) \in \mathbb{R}$ outputs the confidence of r's being matched with t. Suppose that $l \in \{1, 0\}$ is the binary label indicating if (t, r) is a positive or negative region-text pair on x. Our verification model learns to fit the probability for r and t being compatible (a positive pair), i.e., $p(l = 1|x, r, t)$. See Section B in the supplementary materials for a formalized comparison with conditional captioning models.

To this end, we develop a bimodal deep neural network for our model. In particular, $f(x, t, r; \Theta)$ is composed of two single-modality pathways followed by a discriminative pathway. The image pathway $\phi_{\mathrm{rgn}}(x, r; \Theta_{\mathrm{rgn}})$ extracts the d_{rgn}-dim visual representation on the image region r on x. The language pathway $\phi_{\mathrm{txt}}(t; \Theta_{\mathrm{txt}})$ extracts the d_{txt}-dim textual representation for the phrase t. The discriminative pathway with parameters Θ_{dis} dynamically generates a classifier for visual representation according to the textual representation, and predicts if r and t are matched on x. The full model is specified by $\Theta = (\Theta_{\mathrm{txt}}, \Theta_{\mathrm{rgn}}, \Theta_{\mathrm{dis}})$.

3.2. Visual and linguistic pathways

RoI-pooling image network. We suppose the regions of interest are given by an existing region proposal method (e.g., EdgeBox [62], RPN [46]). We calculate visual representations for all image regions in one pass using the fast R-CNN RoI-pooling pipeline. State-of-the-art image classification networks, including the 16-layer VGGNet [49] and ResNet-101 [17], are used as backbone architectures.

Character-level textual network. For an English text phrase t, we encode each of its characters into a 74-dim one-hot vector, where the alphabet is composed of 74 printable characters including punctuations and the space. Thus, the t is encoded as a 74-channel sequence by stacking all character encodings. We use a character-level deep CNN [60] to obtain the high-level textual representation of t. In particular, our network has 6 convolutional layers interleav-

ing with 3 max-pooling layers and followed by 2 fully connected layers (see Section A in the supplementary materials for more details). It takes a sequence of a fixed length as the input and produces textual representations of a fixed dimension. The input length is set to be long enough (here, 256 characters) to cover possible text phrases.[2] To avoid empty tailing characters in the input, we replicate the text phrase until reaching the input length limit.

We empirically found that the very sparse input can easily lead to over-sparse intermediate activations, which can create a large portion of "dead" ReLUs and finally result in a degenerate solution. To avoid this problem, we adopt the Leaky ReLU (LReLU) [35] to keep all hidden units active in the character-level CNN.

Other text embedding methods [29, 24, 27] also can be used in the DBNet framework. We use the character-level CNN because of its simplicity and flexibility. Compared to word-based models, it uses lower-dimensional input vectors and has no constraint on the word vocabulary size. Compared to RNNs, it easily allows deeper architectures.

3.3. Discriminative pathway

The discriminative pathway first forms a linear classifier using the textual representation of the phrase t. Its linear combination weights and bias are

$$\mathbf{w}(t) = \mathbf{A}_{\mathbf{w}}^{\top} \boldsymbol{\phi}_{\mathrm{txt}}(t; \Theta_{\mathrm{txt}}), \tag{1}$$

$$b(t) = \mathbf{a}_b^{\top} \boldsymbol{\phi}_{\mathrm{txt}}(t; \Theta_{\mathrm{txt}}), \tag{2}$$

where $\mathbf{A}_{\mathbf{w}} \in \mathbb{R}^{d_{\mathrm{txt}} \times d_{\mathrm{rgn}}}$, $\mathbf{a}_{\mathbf{b}} \in \mathbb{R}^{d_{\mathrm{txt}}}$, and $\Theta_{\mathrm{dis}} = (\mathbf{A}_{\mathbf{w}}, \mathbf{a}_b)$. This classifier is applied to the visual representation of the image region r on x, obtaining the verification confidence predicted by our model:

$$f(x, r, t; \Theta) = \mathbf{w}(t)^{\top} \boldsymbol{\phi}_{\mathrm{rgn}}(x, r; \Theta_{\mathrm{rgn}}) + b(t). \tag{3}$$

Compared to the basic form of the bilinear function $\boldsymbol{\phi}_{\mathrm{txt}}^{\top}(t; \Theta_{\mathrm{txt}}) \mathbf{A}_{\mathbf{w}} \boldsymbol{\phi}_{\mathrm{rgn}}(x, r; \Theta_{\mathrm{rgn}})$, our discriminative pathway includes an additional linear term as the text-dependent bias for the visual representation classifier.

As a natural way for modeling the cross-modality correlation, multiplication is also a source of instability for training. To improve the training stability, we introduce a regularization term $\Gamma_{\mathrm{dynamic}} = \|\mathbf{w}(t)\|_2^2 + |b(t)|^2$ for the dynamic classifier, besides the network weight decay Γ_{decay} for Θ.

4. Model learning

In DBNet, we drive the training of the proposed two-pathway bimodal CNN with a binary classification objective. We pair image regions and text phrases as training samples. We define the ground truth binary label for

---

[2]The Visual Genome dataset has more than 2.8M unique phrases, whose median length in character is 29. Less than 500 phrases has more than 100 characters.
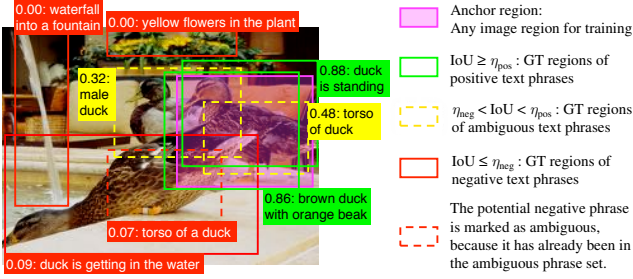
Figure 2: Ground truth labels for region-text pairs (given an arbitrary image region). Phrases are categorized into positive, ambiguous, and negative sets based on the given region's overlap with ground truth boxes (measured by IoU and displayed as the numbers in front of the text phrases). Ambiguous phrases augmented by text similarity is not shown here (see the video in the supplementary materials for an illustration). For visual clarity, $\eta_{\mathrm{neg}} = 0.3$ and $\eta_{\mathrm{pos}} = 0.7$, which are different from the rest of the paper.

each training region-text pair (Section 4.1), and propose a weighted training loss function (Section 4.2).

Training samples. Given M training images $x_1, x_2, \ldots, x_M$, let $\mathcal{G}_i = \{(r_{ij}, t_{ij})\}_{j=1}^{N_i}$ be the set of ground truth annotations for x_i, where N_i is the number of annotations, r_{ij} is the coordinate of the jth region, and t_{ij} is the text phrase corresponding to r_{ij}. When one region is paired with multiple phrases, we take each pair as a separate entry in $\mathcal{G}_i$.

We denote the set of all regions considered on x_i by $\mathcal{R}_i$, which includes both annotated regions $\bigcup_{j=1}^{N_i}\{r_{ij}\}$ and regions given by proposal methods [54, 62, 46]. We write $\mathcal{T}_i = \bigcup \{t_{ij}\}_{j=1}^{N_i}$ for the set of annotated text phrases on x_i, and $\mathcal{T} = \bigcup_{i=1}^{M} \mathcal{T}_i$ for all training text phrases.

4.1. Ground truth labels

Labeling criterion. We assign each possible training region-text pair with a ground truth label for binary classification. For a region r on the image x_i and a text phrase $t \in \mathcal{T}_i$, we take the largest overlap between r and t's ground truth regions as evidence to determine (r, t)'s label. Let $\mathrm{IoU}(\cdot, \cdot)$ denote the intersection over union. The *largest overlap* is defined as

$$\nu_i(r, t) = \max_{r' \in \mathcal{R}_i} \{\mathrm{IoU}(r', r) : (r', t) \in \mathcal{G}_i\}. \qquad (4)$$

In object detection on a limited number of categories (i.e., $\mathcal{T}_i$ consists of category labels), $\nu_i(r, t)$ is usually reliable enough for assigning binary training labels, given the (almost) complete ground truth annotations for all categories.

In contrast, text phrase annotations are inevitably incomplete in the training set. One image region can have an intractable number of valid textual descriptions, including different points of focus and paraphrases of the same description, so annotating all of them is infeasible. Consequently, $\nu_i(r, t)$ cannot always reflect the consistency be-

tween an image region and a text phrase. To obtain reliable training labels, we define positive labels in a conservative manner; and then, we combine text similarity together with spatial IoU to establish the ambiguous text phrase set that reflects potential "false negative" labels. We provide detailed definitions below.

Positive phrases. For a region r on x_i, its positive text phrases (i.e., phrases assigned with positive labels) constitute the set

$$\mathcal{P}_i(r) = \{t \in \mathcal{T}_i : \nu_i(r, t) \geq \eta_{\mathrm{pos}}\}, \qquad (5)$$

where η_{pos} is a high enough IoU threshold ($= 0.9$) to determine positive labels. Some positive phrases may be missing due to incomplete annotations. However, we do not try to recover them (e.g., using text similarity), as "false positive" training labels may be introduced by doing so.

Ambiguous phrases. Still for the region r, we collect the text phrases whose ground truth regions have moderate (neither too large nor too small) overlap with r into a set

$$\mathcal{U}_i(r) = \{t \in \mathcal{T}_i : \eta_{\mathrm{neg}} < \nu_i(r, t) < \eta_{\mathrm{pos}}\}, \qquad (6)$$

where η_{neg} is the IoU lower bound ($= 0.1$). When r's largest IoU with the ground truths of a phrase t lies in $(\eta_{\mathrm{neg}}, \eta_{\mathrm{pos}})$, it is uncertain whether t is positive or negative. In other words, t is *ambiguous* with respect to the region r.

Note that $\mathcal{U}_i(r)$ only contains phrases from $\mathcal{T}_i$. To cover all possible ambiguous phrases from the full set $\mathcal{T}$, we use a text similarity measurement $\mathrm{sim}(\cdot, \cdot)$ to augment $\mathcal{U}_i(r)$ to the finalized ambiguous phrase set

$$\mathcal{A}_i(r) = \{t \in \mathcal{T} : \exists t' \in \mathcal{U}_i(r), \mathrm{sim}(t, t') > \tau\} \backslash \mathcal{P}_i(r), \qquad (7)$$

where we use the METEOR [4] similarity for $\mathrm{sim}(\cdot, \cdot)$ and set the text similarity threshold $\tau = 0.3$.[3]

Labels for region-text pairs. For any image region r on x_i and any phrase $t \in \mathcal{T}$, the ground truth label of (r, t) is

$$y_i(r, t) = \begin{cases} 1, & t \in \mathcal{P}_i(r), \\ \langle \texttt{uncertain} \rangle, & t \in \mathcal{A}_i(r), \\ 0, & \text{otherwise,} \end{cases} \qquad (8)$$

where the pairs of a region and its ambiguous text phrases are assigned with the "uncertain" label to avoid false negative labels. Figure 2 illustrates the region-text label for an arbitrary training image region.

4.2. Weighted training loss

Effective training sets. On the image x_i, the effective set of training region-text pairs is

$$\mathcal{S}_i = \{(r, t) \in \mathcal{R}_i \times \mathcal{T} : y_i(r, t) \neq \langle \texttt{uncertain} \rangle\}, \qquad (9)$$

---

[3] If the METEOR similarity of two phrases is greater than 0.3, they are usually very similar. In Visual Genome, ∼0.25% of all possible pairs formed by the text phrases that occur ≥20 times can pass this threshold.

where, as previously defined, $\mathcal{R}_i$ consists of annotated and proposed regions, and $\mathcal{T}$ consists of all phrases from the training set. We exclude samples of uncertain labels.

We partition $\mathcal{S}_i$ into three subsets according to the value of $y_i(r,t)$ and the origin of the phrase t: $\mathcal{S}_i^{\text{pos}}$ for $y_i(r,t) = 1$, $\mathcal{S}_i^{\text{neg}}$ for $y_i(r,t) = 0 \wedge t \in \mathcal{T}_i$, and $\mathcal{S}_i^{\text{rest}}$ for all negative region-text pairs containing phrases from the rest of the training set (i.e., not from x_i).

Per-image training loss Let $f_i(r,t) = f(x_i, r, t; \Theta) \in \mathbb{R}$ for notation convenience; and, let $\ell(\cdot, \cdot)$ be a binary classification loss, in particular, the cross-entropy loss of logistic regression. We define the training loss on x_i as the summation of three parts:

$$L_i = \lambda_{\text{pos}} L_i^{\text{pos}} + \lambda_{\text{neg}} L_i^{\text{neg}} + \lambda_{\text{rest}} L_i^{\text{rest}}, \qquad (10)$$

$$L_i^{\text{pos}} = \frac{1}{|\mathcal{S}_i^{\text{pos}}|} \sum_{(r,t) \in \mathcal{S}_i^{\text{pos}}} \ell\left(f_i(r,t), 1\right), \qquad (11)$$

$$L_i^{\text{neg}} = \frac{1}{|\mathcal{S}_i^{\text{neg}}|} \sum_{(r,t) \in \mathcal{S}_i^{\text{neg}}} \ell\left(f_i(r,t), 0\right), \qquad (12)$$

$$L_i^{\text{rest}} = \frac{\sum_{(r,t) \in \mathcal{S}_i^{\text{rest}}} \text{freq}(t) \cdot \ell\left(f_i(r,t), 0\right)}{\sum_{(r,t) \in \mathcal{S}_i^{\text{rest}}} \text{freq}(t)}, \qquad (13)$$

where $\text{freq}(t)$ is t's frequency of occurrences in the training set. We normalize and re-weight the loss for each of the three subsets of $\mathcal{S}_i$ separately. In particular, we set $\lambda_{\text{pos}} = \lambda_{\text{neg}} + \lambda_{\text{rest}} = 1$ to balance the positive and negative training loss. The values of λ_{neg} and λ_{rest} are implicitly determined by the numbers of text phrases that we choose inside and outside x_i during stochastic optimization.

The training loss functions in most existing work on natural-language visual localization [21, 23] use only positive samples for training, which is similar to solely using L_i^{pos}. The method in [38] also considers the negative case (similar to L_i^{neg}), but it is less flexible and not extensible to the case of L_i^{rest}. The recurrent neural language model can encourage a certain amount of discriminativeness on word selection, but not on entire text phrases as ours.

Full training objective. Summing up the training loss for all images together with weight decay for the whole neural network and the regularization for the text-specific dynamic classifier (Section 3.3), the full training objective is:

$$\min_{\Theta} \frac{1}{M} \sum_{i=1}^{M} L_i + \beta_1 \Gamma_{\text{decay}} + \beta_2 \Gamma_{\text{dynamic}}, \qquad (14)$$

where we set $\beta_1 = 5 \times 10^{-4}$ and $\beta_2 = 10^{-8}$. Model optimization is in Section C of the supplementary materials.

5. Experiments

Dataset. We evaluated the proposed DBNet on the Visual Genome dataset [30]. It contains 108,077 images, where ~5M regions are annotated with text phrases in order to densely cover a wide range of visual entities.

We split the Visual Genome datasets in the same way as in [23]: 77,398 images for training, 5,000 for validation (tuning model parameters), and 5000 for testing; the remaining 20,679 images were not included (following [23]).

The text phrases were annotated from crowd sourcing and included a significant portion of misspelled words. We corrected misspelled words using the Enchant spell checker [1] from AbiWord. After that, there were 2,113,688 unique phrases in the training set and 180,363 unique phrases in the testing set. In the test set, about one third (61,048) of the phrases appeared in the training set, and the remaining two thirds (119,315) were unseen. About 43 unique phrases were annotated with ground truth regions per image. All experimental results are reported on this dataset.

Models. We constructed the fast R-CNN [13]-style visual pathway of DBNet based on either the 16-layer VGGNet (Model-D in [49]) or ResNet-101 [17]. In most experiments, we used VGGNet for fair comparison with existing works (which also use VGGNet) and less evaluation time. ResNet-101 was used to further improve the accuracy.

We compared DBNet with two image captioning based localization models: DenseCap [23] and SCRC [21]. In DBNet, the visual pathway was pretrained for object detection using the faster R-CNN [46] on the PASCAL VOC 2007+2012 trainval set [10]. The linguistic pathway was randomly initialized. Pretrained VGGNet on ImageNet ILSVRC classification dataset [8] was used to initialize DenseCap, and the model was trained to match the dense captioning accuracy reported by Johnson et al. [23]. We found that the faster R-CNN pretraining did not benefit DenseCap (see Section E.1 of the supplementary materials). The SCRC model was additionally pretrained for image captioning on MS COCO [33] in the same way as Hu et al. [21] did.

We trained all models using the training set on Visual Genome and evaluated them for both localization on single images and detection on multiple images. We also assessed the usefulness of the major components of our DBNet.

5.1. Single image localization

In the localization task, we took all ground truth text phrases annotated on an image as queries to localize the associated objects by maximizing the network response over proposed image regions.

Evaluation metrics. We used the same region proposal method to propose bounding boxes for all models, and we used the non-maximum suppression (NMS) with the IoU threshold 0.3 to localize a few boxes. The performance was evaluated by the recall of ground truth regions of the query phrase (see Section D of the supplementary materials for

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		
DC-RPN 500	16-layer VGGNet	DenseCap	52.5	38.9	27.0	17.1	9.5	4.3	1.5	0.117	0.184
		DBNet	**57.4**	**46.9**	**37.8**	**29.4**	**21.3**	**13.6**	**7.0**	**0.168**	**0.250**
EdgeBox 500	16-layer VGGNet	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 w/o bias term	52.3	43.8	36.3	29.3	22.4	15.7	9.4	0.124	0.246
		DBNet w/o VOC pretraining	54.3	45.0	36.6	28.8	21.3	14.4	8.2	0.144	0.245
		DBNet	**54.8**	**45.9**	**38.3**	**30.9**	**23.7**	**16.6**	**9.9**	**0.152**	**0.258**
	ResNet-101	DBNet	**59.6**	**50.5**	**42.3**	**34.3**	**26.4**	**18.6**	**11.2**	**0.205**	**0.284**

Table 1: 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.

DenseCap performance	Recall / % for IoU@			Median IoU
	0.1	0.3	0.5	
Small test set in [23]	56.0	34.5	15.3	0.137
Test set in this paper	50.5	24.7	8.1	0.103

Table 2: Localization accuracy of DenseCap on the small test set (1000 images and 100 test queries) used in [23] and the full test set (5000 images and >0.2M queries) used in this paper. 1000 boxes (at most) per image are proposed using the DenseCap RPN.

(a) IoU@0.5   **(b)** IoU@0.7

Figure 3: Top-k localization recall under two overlapping thresholds. VGGNet and EdgeBox 500 are used in all methods.

a discussion on recall and precision for localization tasks). If one of the proposed bounding boxes with the top-k network responses had a large enough overlap (determined by an IoU threshold) with the ground truth bounding box, we took it as a successful localization. If multiple ground truth boxes were on the same image, we only required the localized boxes to match one of them. The final recall was averaged over all test cases, i.e., per image and text phrase. Median and mean overlap (IoU) between the top-1 localized box and the ground truth were also considered.

DBNet outperforms captioning models. We summarize the top-1 localization performance of different methods in Table 1, where 500 bounding boxes were proposed for testing. DBNet outperforms DenseCap and SCRC under all metrics. In particular, DBNet's recall was more than twice as high as the other two methods for the IoU threshold at 0.5 (commonly used for object detection [10, 33]) and about 4 times higher for IoU at 0.7 (for high-precision localization [12, 61]).

Johnson et al. [23] reported DenseCap's localization accuracy on a much smaller test set (1000 images and 100 test queries in total), which is not comparable to our exhaustive test settings (Table 2 for comparison). We also note that different region proposal methods (EdgeBox and DenseCap RPN) did not make a big difference on the localization performance. We used EdgeBox for the rest of our evaluation.

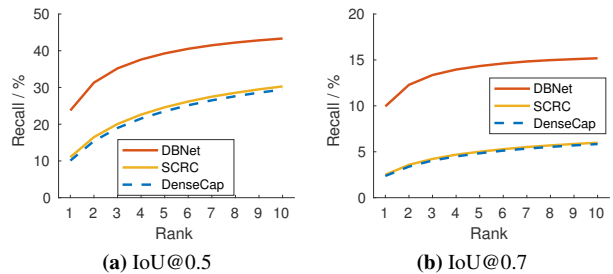Figure 3 shows the top-k recall ($k = 1, 2, \ldots, 10$) in curves. SCRC is slightly better than DenseCap, possibly due to the global context features used in SCRC. DBNet outperforms both consistently with a significant margin, thanks to the effectiveness of discriminative training.

Dynamic bias term improves performance. The text-dependent bias term introduced in (2) and (3) makes our method for fusing visual and linguistic representations different from the basic bilinear functions (e.g., used in [44]) and more similar to a visual feature classifier. As in Table 1, this dynamic bias term led to $> 20\%$ relative improvement on median IoU and $\sim 5\%$ ($2.5\% \sim 0.5\%$ absolute) relative improvement on recall at all IoU thresholds.

Transferring knowledge benefits localization accuracy. Pretraining the visual pathway of DBNet for object detection on PASCAL VOC showed minor benefit on recall at lower IoU thresholds, but it brought 10% and 17% relative improvement to the recall for the IoU threshold at 0.5 and 0.7, respectively. See Section E.1 in the supplementary materials for more results, where we showed that DenseCap did not get benefit from the same technique.

Qualitative results. We visually compared the localization results of DBNet and DenseCap in Figure 4. In many cases, DBNet localized the queried entities at more reasonable locations. More examples are provided in Section F of the supplementary materials.

More quantitative results. In the supplementary materials, we studied the performance improvement of the learned
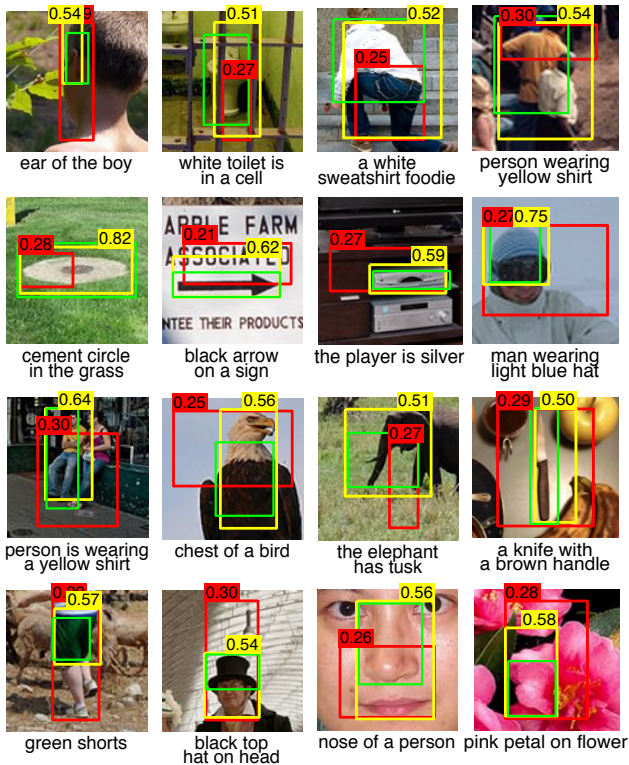
Figure 4: Qualitative comparison between DBNet and Dense-Cap on localization task. Green boxes: ground truth; Red boxes: DenseCap; Yellow boxes: DBNet.

Average precision / %	IoU@0.3		IoU@0.5		IoU@0.7	
	mAP	gAP	mAP	gAP	mAP	gAP
DenseCap	36.2	1.8	15.7	0.5	3.4	0.0
SCRC	38.5	2.2	16.5	0.5	3.4	0.0
DBNet	**48.1**	**23.1**	**30.0**	**10.8**	**11.6**	**2.1**
DBNet w/ Res	51.1	24.2	32.6	11.5	12.9	2.2

(a) **Level-0:** Only positive images per text phrase.

Average precision / %	IoU@0.3		IoU@0.5		IoU@0.7	
	mAP	gAP	mAP	gAP	mAP	gAP
DenseCap	22.9	1.0	10.0	0.3	2.1	0.0
SCRC	37.5	1.7	16.3	0.4	3.4	0.0
DBNet	**45.5**	**21.0**	**28.8**	**9.9**	**11.4**	**2.0**
DBNet w/ Res	48.3	22.2	31.2	10.7	12.6	2.1

(b) **Level-1:** The ratio between the positive and negative images is 1:1 per text phrase.

Average precision / %	IoU@0.3		IoU@0.5		IoU@0.7	
	mAP	gAP	mAP	gAP	mAP	gAP
DenseCap	4.1	0.1	1.7	0.0	0.3	0.0
DBNet	**26.7**	**8.0**	**17.7**	**3.9**	**7.6**	**0.9**
DBNet w/ Res	29.7	9.0	19.8	4.3	8.5	0.9

(c) **Level-2:** The ratio between the positive and negative images is at least 1:5 (minimum 20 negative images and 1:5 otherwise) per text phrase.

Table 3: Detection average precision using query set of three levels of difficulties. mAP: mean AP over all text phrases. gAP: AP over all test cases. VGGNet is the default visual CNN for all methods. "DBNet w/ Res" denotes our DBNet with ResNet-101.

models over random guessing and the upper bound performance due to the limitation of region proposal methods (Section E.2). We also evaluated DBNet using queries in a constrained form (Section E.3), where the high query complexity was demonstrated as a significant source of failures for natural language visual localization.

5.2. Detection on multiple images

In the detection task, the model needs to verify the existence and quantity of queried visual entities in addition to localizing them, if any. Text phrases not associated with any image regions can exist in the query set of an image, and evaluation metrics can be defined by extending those used in traditional object detection.

Query sets. Due to the huge total number of possible query phrases, it is practical to test only a subset of phrases on a test image. We developed query sets in three difficulty levels $(0, 1, 2)$. For a text phrase, a test image is *positive* if at least one ground truth region exists for the phrase; otherwise, the image is *negative*.

- *Level-0:* The query set was the same as in the localization task, so every text phrase was tested only on its positive images ($\sim$43 phrases per image).
- *Level-1:* For each text phrase, we randomly chose the

same number of negative images and the positive images ($\sim$92 phrases per image).
- *Level-2:* The number of negative images was either 5 times the number of positive images or 20 (whichever was larger) for each test phrase ($\sim$775 phrases per image). This set included relatively more negative images (compared to positive images) for infrequent phrases.

As the level went up, it became more challenging for a detector to maintain its precision, as more negative test cases are included. In the level-1 and level-2 sets, text phrases depicting obvious non-object "stuff", such as sky, were removed to better fit the detection task. Then, 176,794 phrases (59,303 seen and 117,491 unseen) remained.

Evaluation metrics. We measured the detection performance by average precision (AP). In particular, we computed AP independently for each query phrase (comparable to a category in traditional object detection [10]) over its test images, and reported the *mean AP* (**mAP**) over all query phrases. Like traditional object detection, the score threshold for a detected region is category/phrase-specific.

For more practical natural-language visual detection, where the query text may not be known in advance, we also directly computed AP over all test cases. We term it *global*

AP (**gAP**), which implies a universal decision threshold for any query phrase. Table 3 summarizes mAPs and gAPs under different overlapping thresholds for all models.

DBNet shows higher per-phrase performance. DBNet achieved consistently stronger performance than DenseCap and SCRC in terms of mAP, indicating that DBNet produced more accurate detection per given phrase. Even for the challenging IoU threshold of 0.7, DBNet still showed reasonable performance. The mAP results suggest the effectiveness of discriminative training.

DBNet scores are better "calibrated". Achieving good performance in gAP is challenging as it assumes a phrase-agnostic, universal decision threshold. For IoU at 0.3 and 0.5, DenseCap and SCRC showed very low performance in terms of gAP, and DBNet dramatically ($10 \sim 20\times$) outperformed them. For IoU at 0.7, DenseCap and SCRC were unsuccessful, while DBNet could produce a certain degree of positive results. The gAP results suggest that the responses of DBNet are much better calibrated among different text phrases than captioning models, supporting our hypothesis that distributions on a binary decision space are easier to model than those on the huge natural language space.

Robustness to negative and rare cases. The performance of all models dropped as the query set became more difficult. SCRC appeared to be more robust than DenseCap for negative test cases (level-1 performance). DBNet showed superior performance in all difficulty levels. Particularly for the level-2 query set, DenseCap's performance dropped significantly compared to the level-1 case, which suggests that it probably failed at handling rare phrases (note that relatively more negative images are included in the level-2 set for rare phrases). For IoU at 0.5 and 0.7, DBNet's level-2 performance was even better than the level-0 performance of DenseCap and SCRC. We did not test SCRC on the level-2 query set because of its high time consumption.[4]

Qualitative results. We showed qualitative results of DBNet detection on selected examples in Figure 5. More comprehensive (random and failed) examples are provided in Section G of the supplementary materials. Our DBNet could detect diverse visual entities, including objects with attributes (e.g., "a bright colored snow board"), objects in context (e.g., "little boy sitting up in bed"), object parts (e.g., "front wheel of a bicycle"), and groups of objects (e.g.,"bikers riding in a bicycle lane").

5.3. Ablation study on training strategy

We did ablation studies for three components of our DB-Net training strategy: 1) pruning ambiguous phrases ($\mathcal{A}_i(r)$

---

[4]For level-2 query set, DBNet and DenseCap cost ∼0.5 min to process one image (775 queries) when using the VGGNet and a Titan X card. SCRC takes nearly 10 minutes with the same setting. In addition, DBNet took 2–3 seconds to process one image when using level-0 query set.

defined in Eq. (7)), 2) training with negative phrases from other images (L_i^{rest}), and 3) finetuning the visual pathway.

As shown in Table 4, the performance of the most basic training strategy is better than DenseCap and SCRC, due to the effectiveness of discriminative training. Ambiguous phrase pruning led to significant performance gain, by improving the correctness of training labels, where no "pruning ambiguous phrases" means setting $\mathcal{A}_i(r) = \emptyset$. More quantitative analysis on tuning the text similarity threshold τ are provided in Section E.4 of the supplementary materials. Inter-image negative phrases did not benefit localization performance, since localization is a single-image task. However, this mechanism improved the detection performance by making the model more robust to diverse negative cases. As expected in most vision tasks, finetuning pretrained classification network boosted the performance of our models. In addition, upgrading the VGGNet-based visual pathway to ResNet-101 led to another clear gain in DBNet's performance (Table 1 and 3).

6. Conclusion

We demonstrated the importance of discriminative learning for natural-language visual localization. We proposed the discriminative bimodal neural network (DBNet) to allow flexible discriminative training objectives. We further developed a comprehensive training strategy to extensively and properly leverage negative observations on training data. DBNet significantly outperformed the previous state-of-the-art based on caption generation models. We also proposed quantitative measurement protocols for natural-language visual detection. DBNet showed more robustness against rare queries compared to existing methods and produced detection scores with better calibration over various text queries. Our method can be potentially improved by combining its discriminative objective with a generative objective, such as image captioning.

Acknowledgements

References

[1] AbiWord. Enchant spell checker. http://www.abisource.com/projects/enchant/. 5

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Kleina. Neural module networks. In *CVPR*, 2016. 1
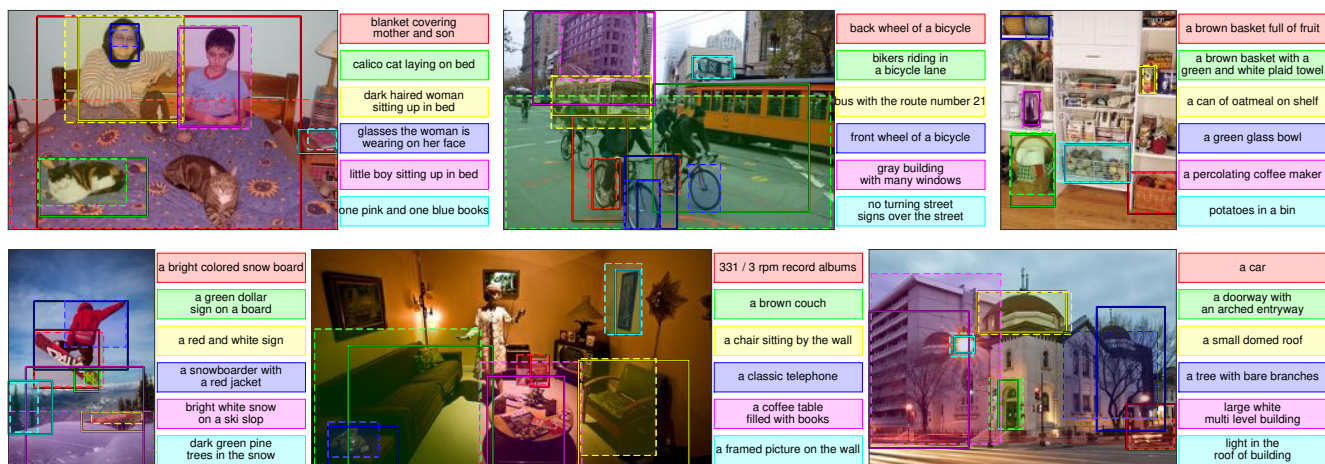
[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra,

Figure 5: Qualitative detection results of DBNet with ResNet-101. We show detection results of six different text phrases on each image. For each image, the colors of bounding boxes correspond to the colors of text tags on the right. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results.

Prune ambiguous phrases	Phrases from other images	Finetune visual pathway	Localization					Detection (Level-1)					
			Recall / % for IoU@			Median	Mean	mAP / % for IoU@			gAP / % for IoU@		
			0.3	0.5	0.7	IoU	IoU	0.3	0.5	0.7	0.3	0.5	0.7
No	No	No	30.6	17.5	7.8	0.066	0.211	35.5	22.0	8.6	8.3	3.1	0.4
Yes	No	No	34.5	21.2	9.0	0.113	0.237	39.0	24.6	9.7	15.5	7.4	1.6
Yes	Yes	No	34.7	21.1	8.8	0.119	0.238	41.3	25.6	10.0	17.2	7.9	1.6
Yes	Yes	Yes	**38.3**	**23.7**	**9.9**	**0.152**	**0.258**	**45.5**	**28.8**	**11.4**	**21.0**	**9.9**	**2.0**

Table 4: Ablation study of DBNet's major components. The visual pathway is based on the 16-layer VGGNet.

C. Lawrence Zitnick, and D. Parikh. VQA: Visual question answering. In *CVPR*, 2015. 1

[4] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005. 4

[5] X. Chen and C. L. Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, 2015. 1

[6] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 3

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5, 14

[9] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):

677–691, April 2017. 1

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 5, 6, 7, 14

[11] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 1, 3

[12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 6

[13] R. Girshick. Fast R-CNN. In *ICCV*, 2015. 2, 5

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1): 142–158, Jan 2016. 1, 2, 3

[15] A. Graves. Generating sequences with recurrent neural networks. *arXiv:1308.0850*, 2013. 1, 2

[16] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014. 2

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3, 5

[18] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 1

[19] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[20] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 2

[21] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *CVPR*, 2016. 1, 2, 5, 13

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014. 13

[23] J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 1, 2, 3, 5, 6, 13

[24] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. In *ACL*, 2014. 2, 3

[25] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1, 2

[26] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1

[27] Y. Kim. Convolutional neural networks for sentence classification. *EMNLP*, 2014. 2, 3

[28] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 13

[29] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler. Skip-thought vectors. In *NIPS*, 2015. 2, 3

[30] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017. 1, 5

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 2

[32] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 13

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*. 5, 6

[34] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detections. In *ECCV*, 2016. 2

[35] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 3

[36] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *CVPR*, 2015. 1

[37] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). In *ICLR*, 2015. 1

[38] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 1, 2, 5, 13

[39] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, 2010. 1, 2

[40] V. Nagaraja, V. Morariu, and L. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 2

[41] H. Noh, P. H. Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*, 2016. 1

[42] W. Ouyang, X. Zeng, X. Wang, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, H. Li, C. C. Loy, K. Wang, J. Yan, and X. Tang. DeepID-Net: Deformable deep convolutional neural networks for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. 2

[43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2

[44] S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. In *IEEE Computer Vision and Pattern Recognition*, 2016. 1, 2, 6

[45] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In *ICML*, 2016. 1

[46] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2, 3, 4, 5, 14

[47] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 2

[48] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated recogni-

tion, localization and detection using convolutional networks. In *ICLR*, 2014. 2

[49] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1, 2, 3, 5

[50] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In *ICML*, 2011. 1, 2

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2

[52] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2, 3

[53] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. 2016. 2

[54] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104 (2):154–171, 2013. 4

[55] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 2

[56] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 1, 2

[57] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 1

[58] L. Yu, P. Poirson, S. Yang, A. Berg, and T. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 2

[59] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 2

[60] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015. 2, 3

[61] Y. Zhang, K. Sohn, R. Villegas, G. Pan, and H. Lee. Improving object detection with deep convolutional networks via bayesian optimization and structured prediction. In *CVPR*, 2015. 2, 6, 14

[62] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 3, 4, 13

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

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 precision = recall/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 proposal	Localization model	Accuracy / % for IoU@							Median IoU	Mean IoU
		0.1	0.2	0.3	0.4	0.5	0.6	0.7		
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 IoU	Mean IoU
	0.1	0.2	0.3	0.4	0.5	0.6	0.7		
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.

See results on the next page.

F.1. More qualitative comparison with DenseCap

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

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.
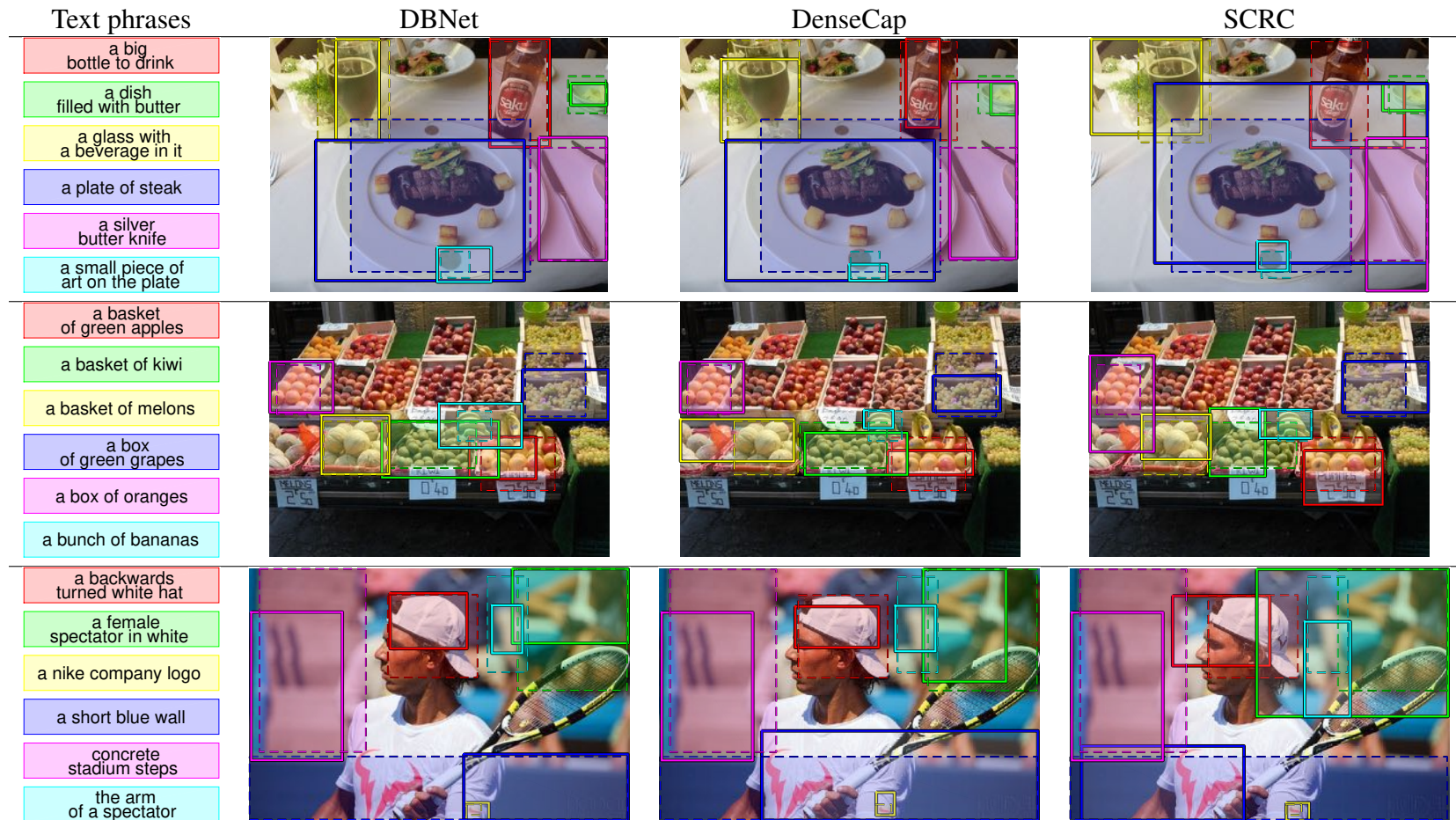
Figure 8: Qualitative detection results of DBNet, DenseCap, and SCRC when the number of ground truth is known. Detection results of six different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.
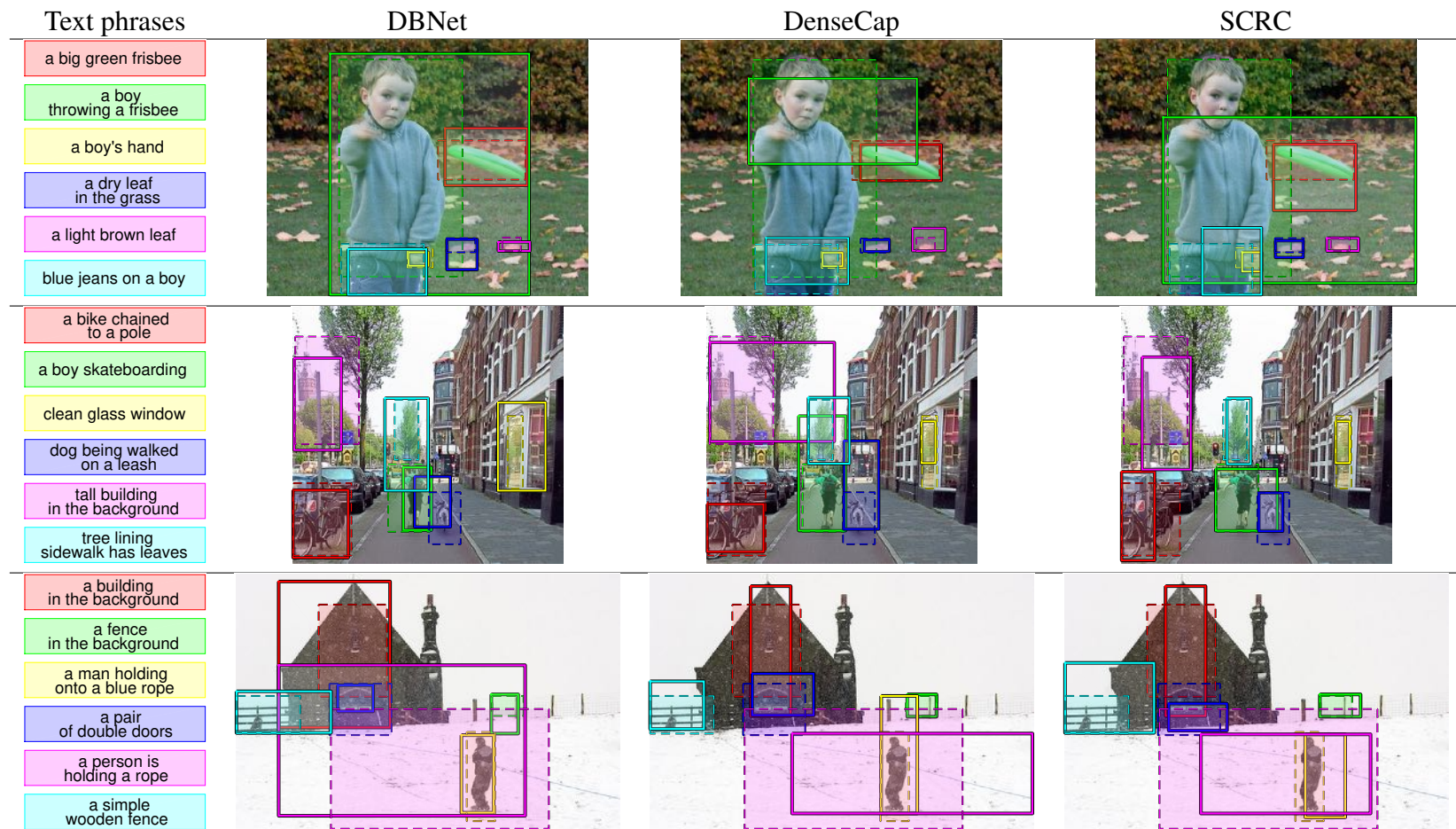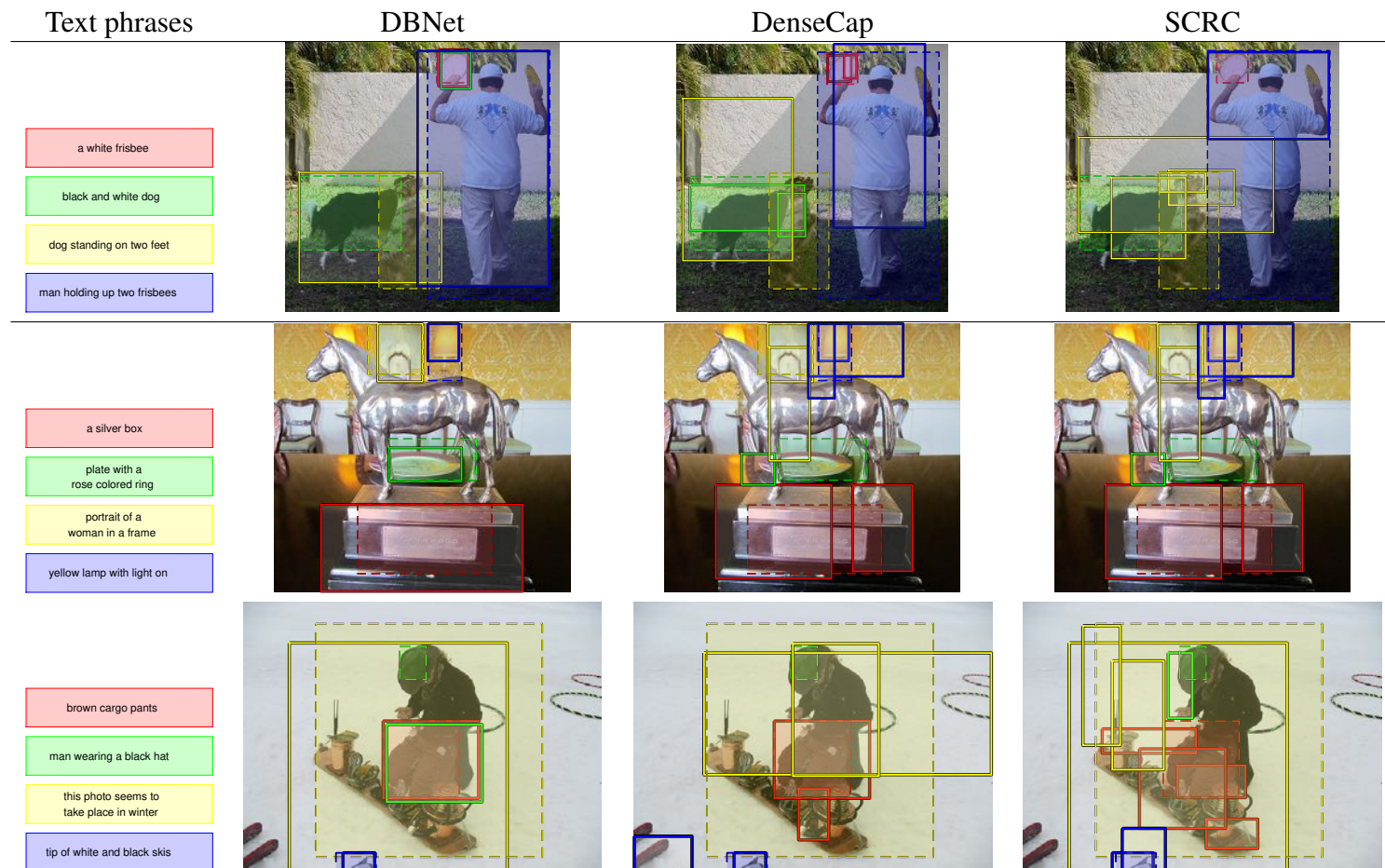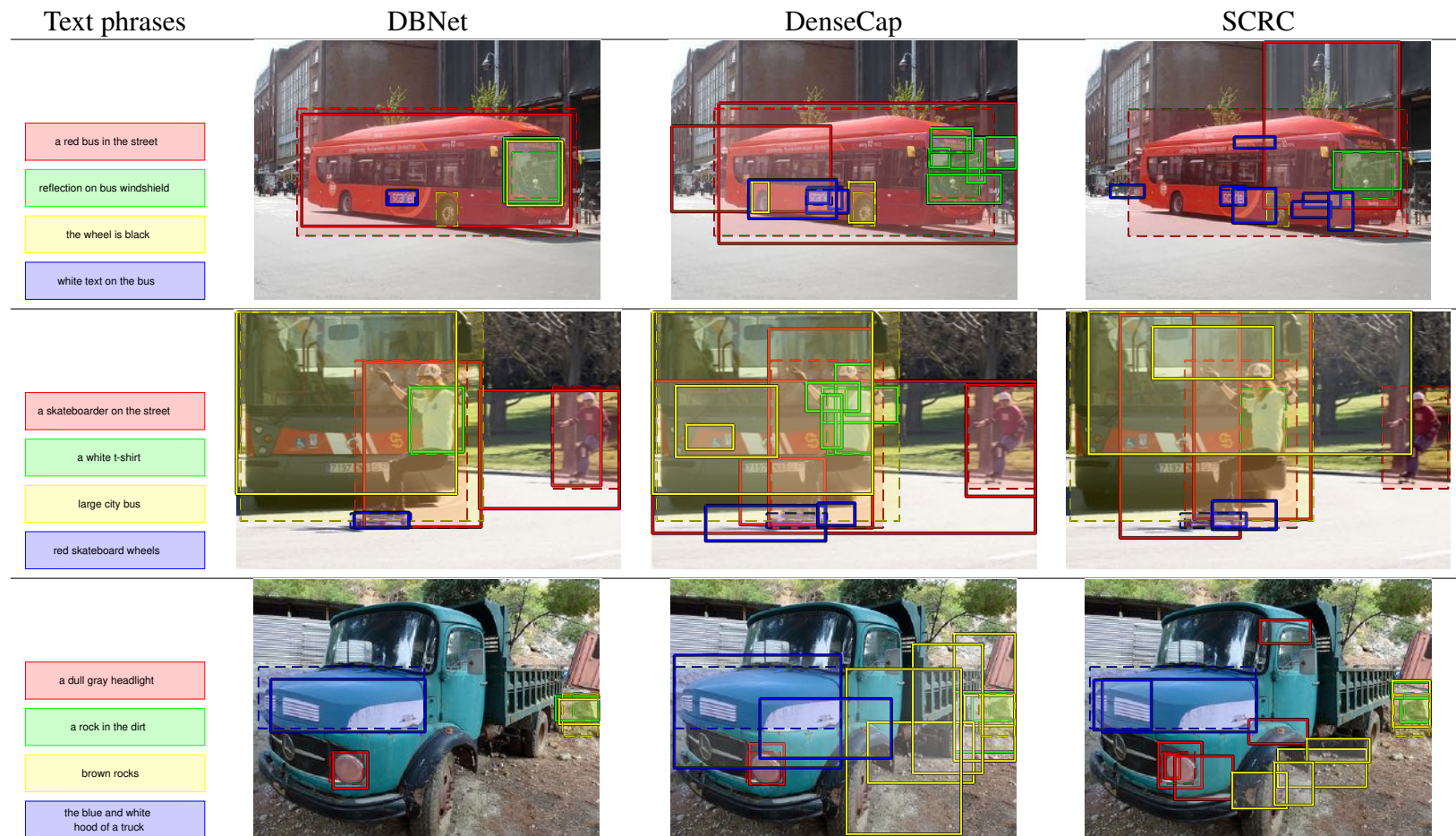
Figure 9: (continued from Figure 8) Qualitative detection results of DBNet, DenseCap, and SCRC when the number of ground truth is known. Detection results of six different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.
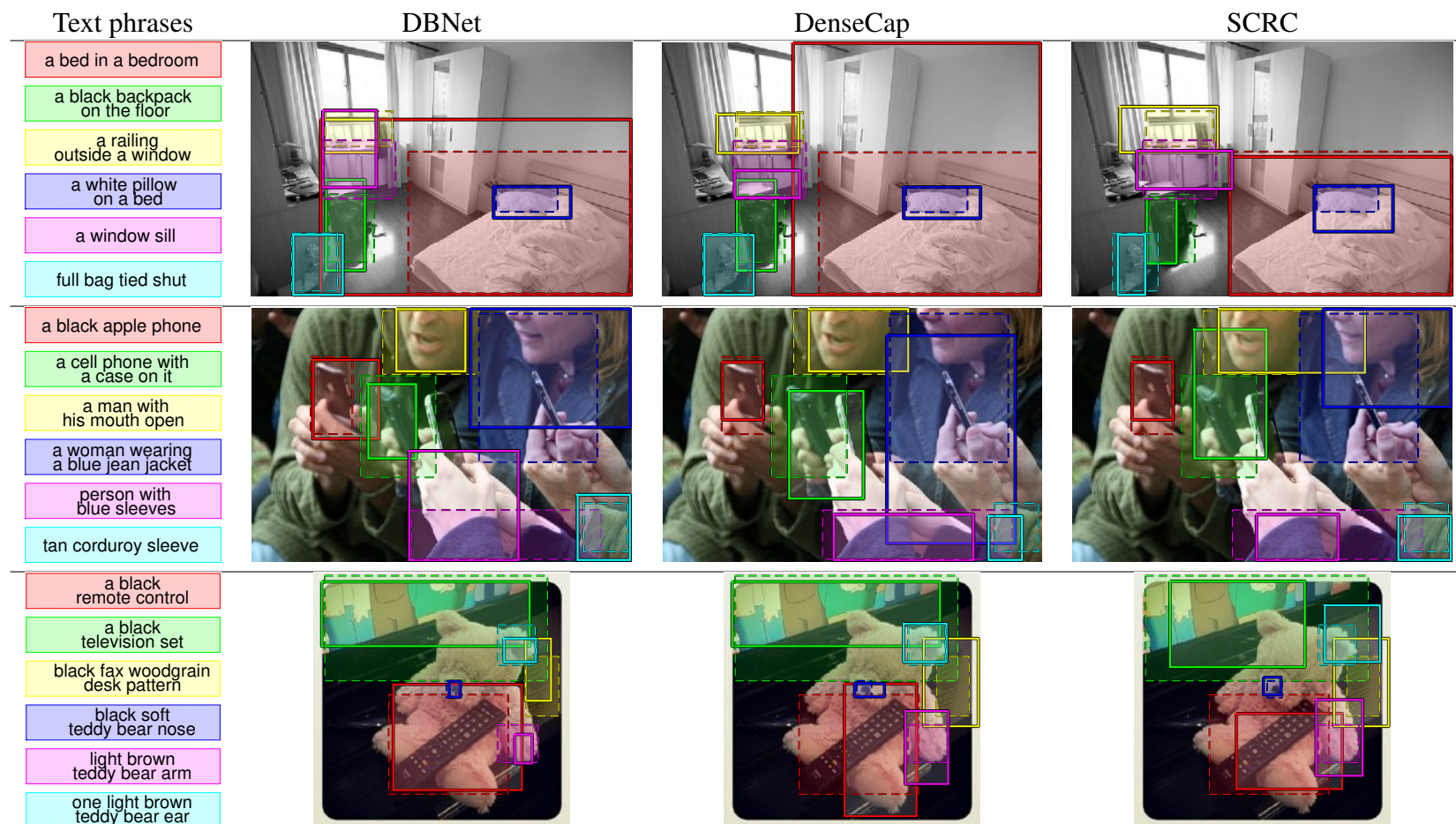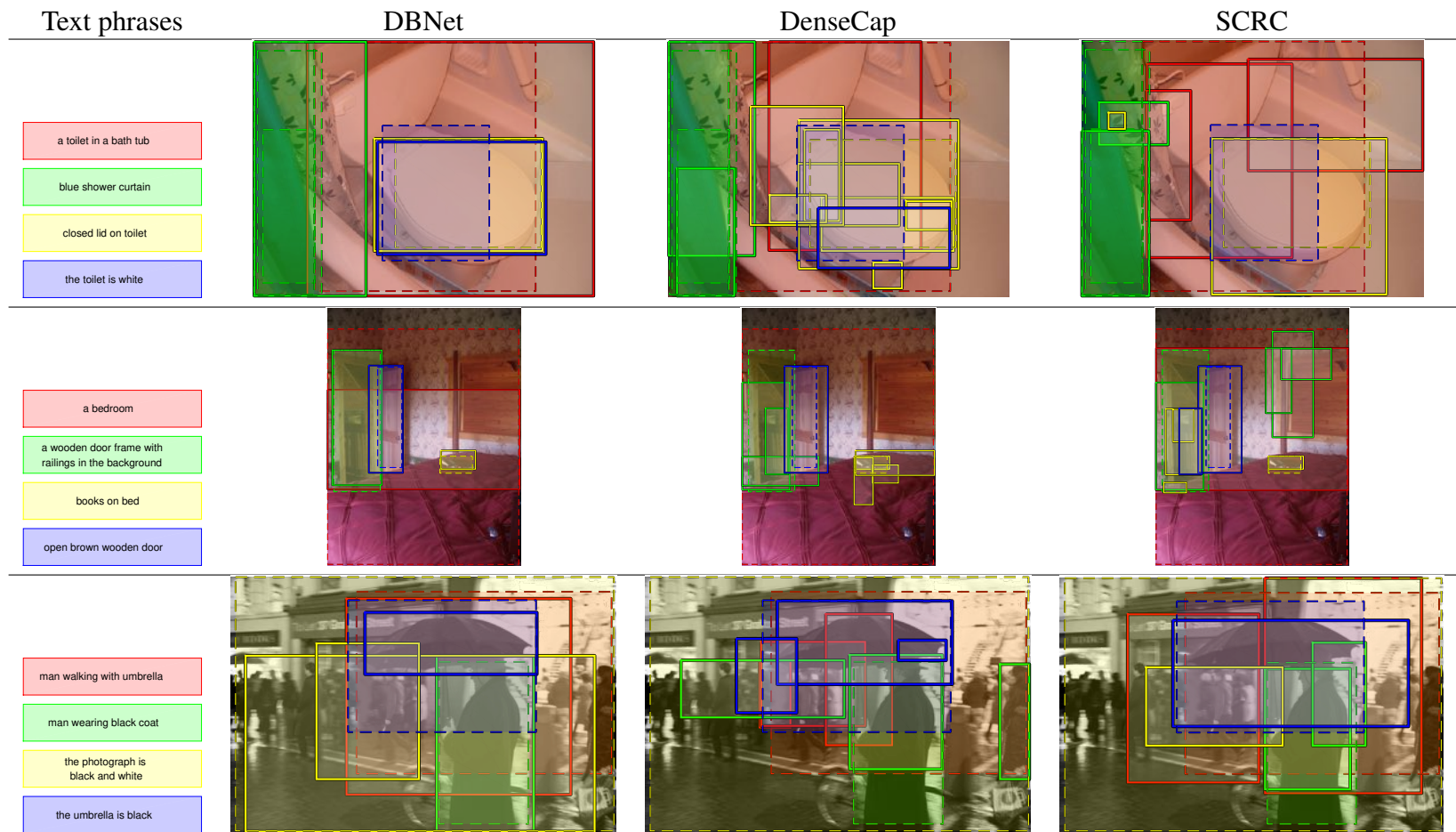
Figure 10: (continued from Figure 9) Qualitative detection results of DBNet, DenseCap, and SCRC when the number of ground truth is known. Detection results of six different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The 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.
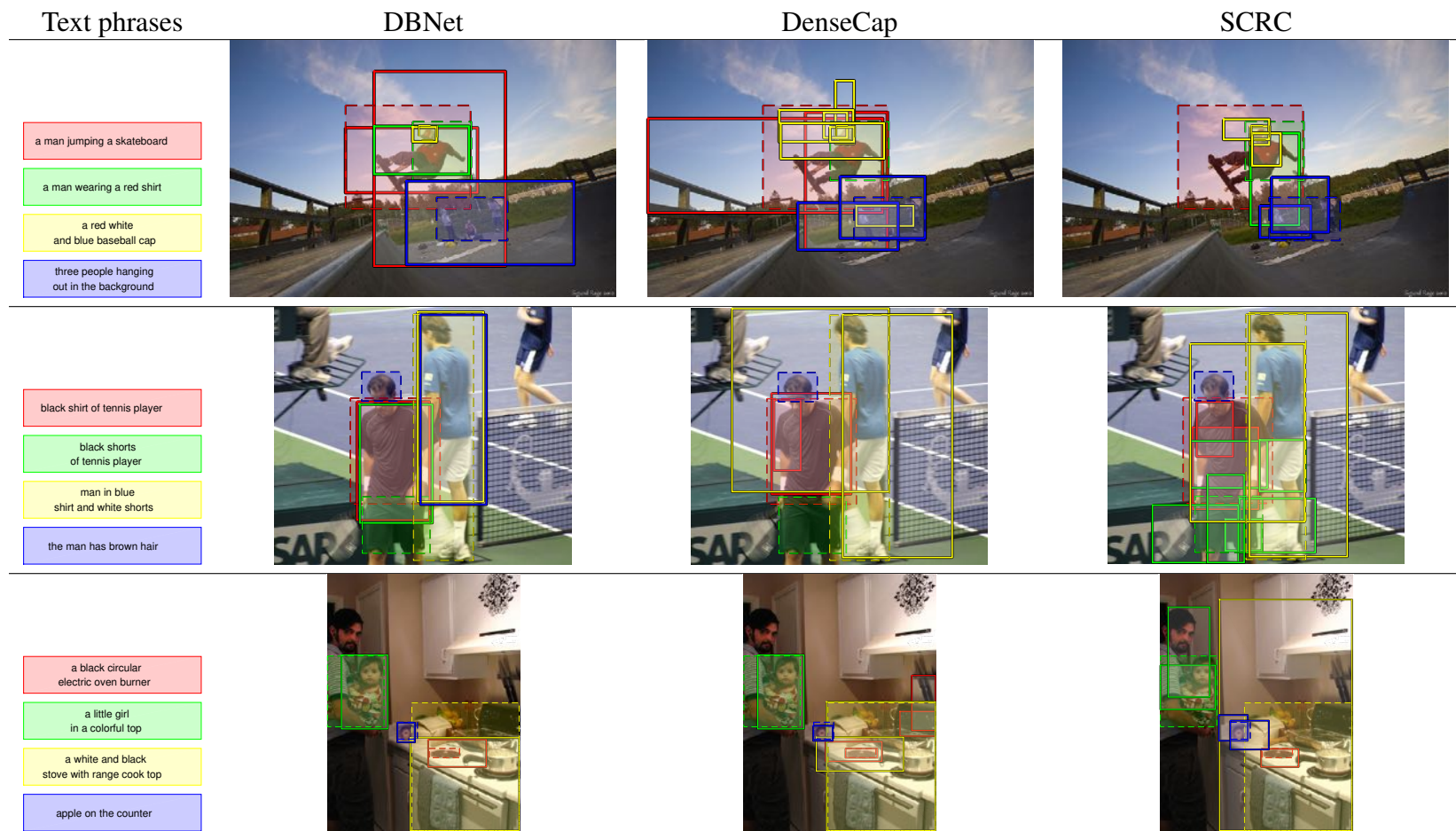
Figure 11: Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.
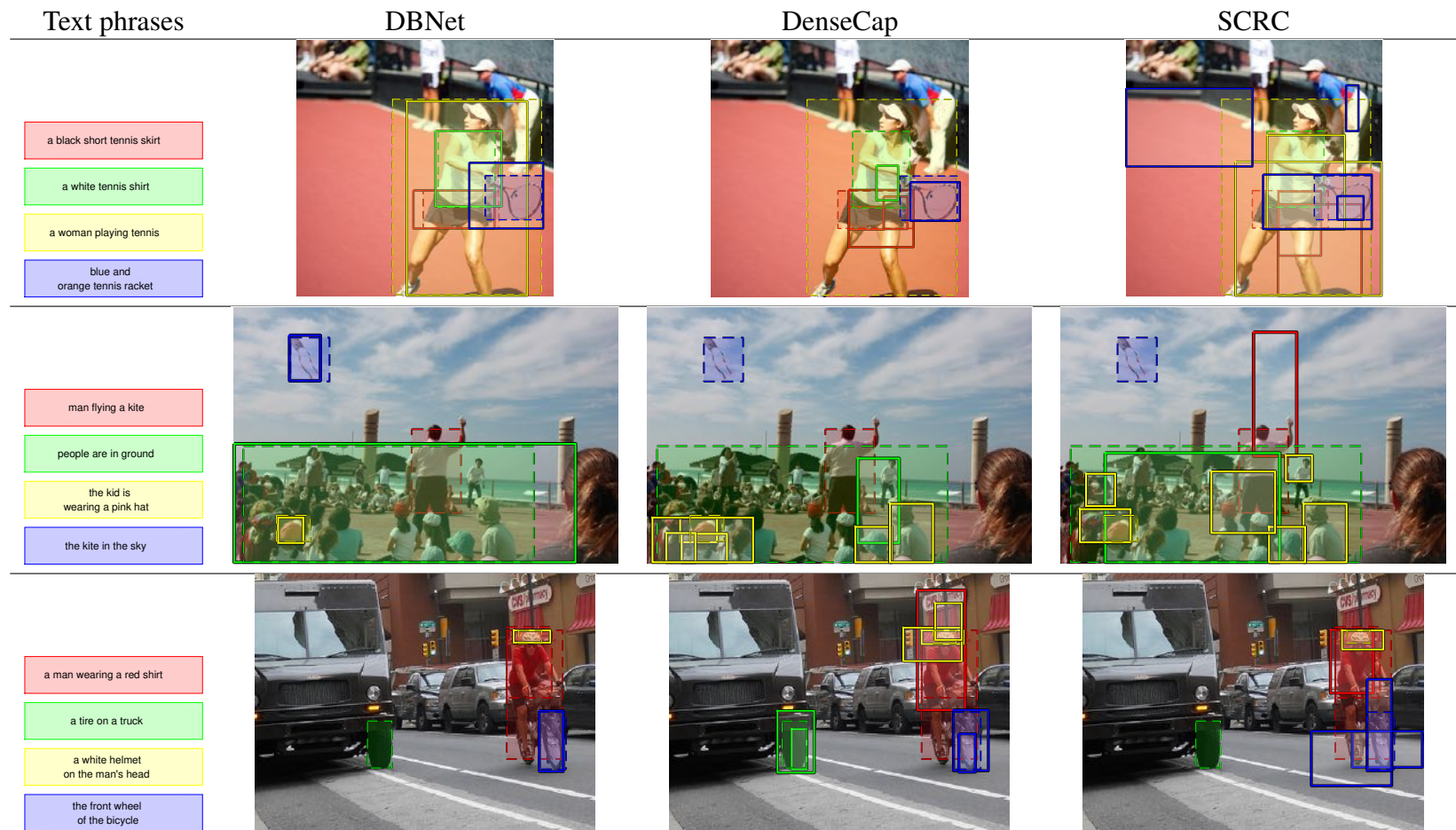
Figure 12: (continued from Figure 11) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.
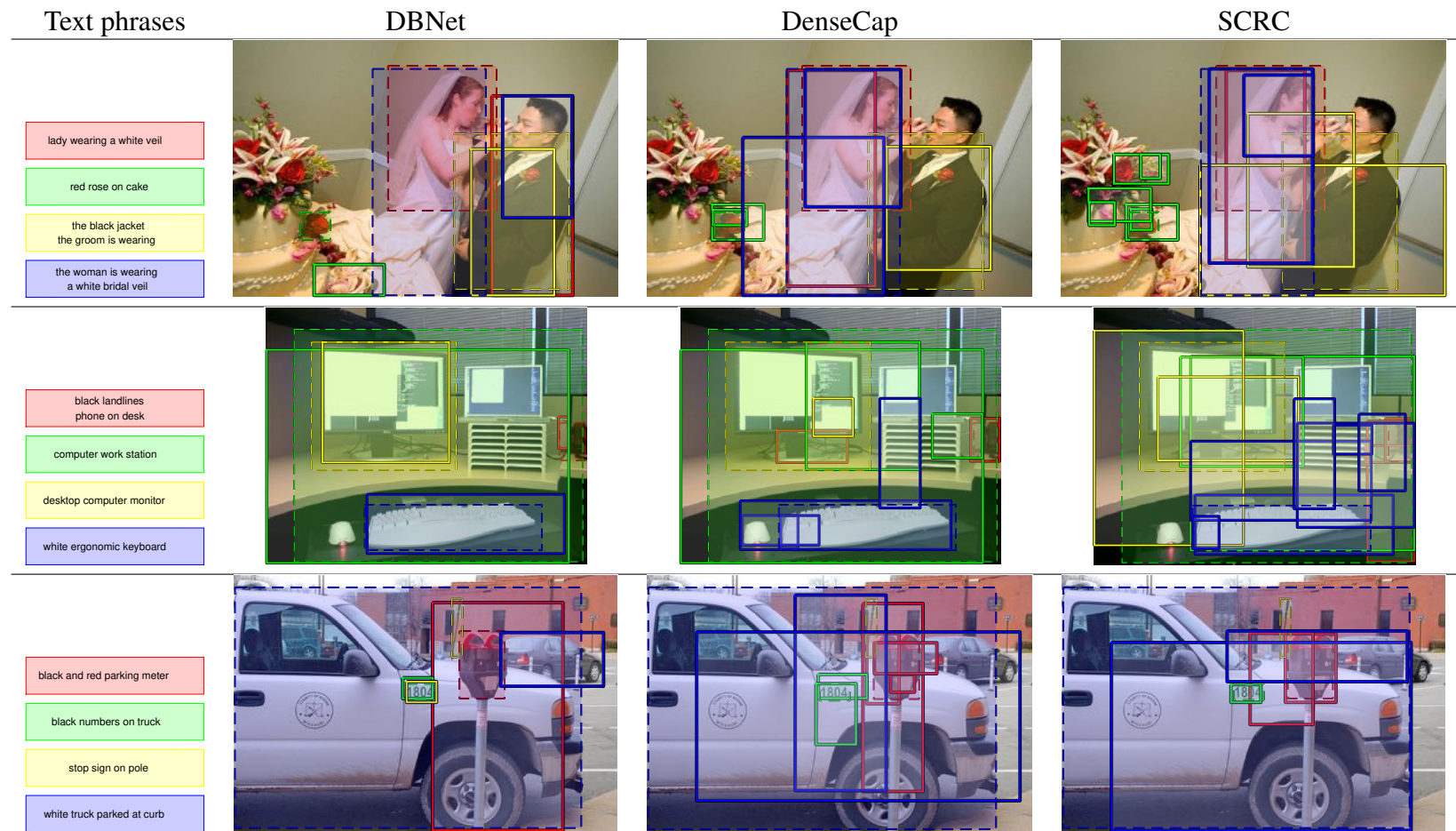
Figure 13: (continued from Figure 12) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.
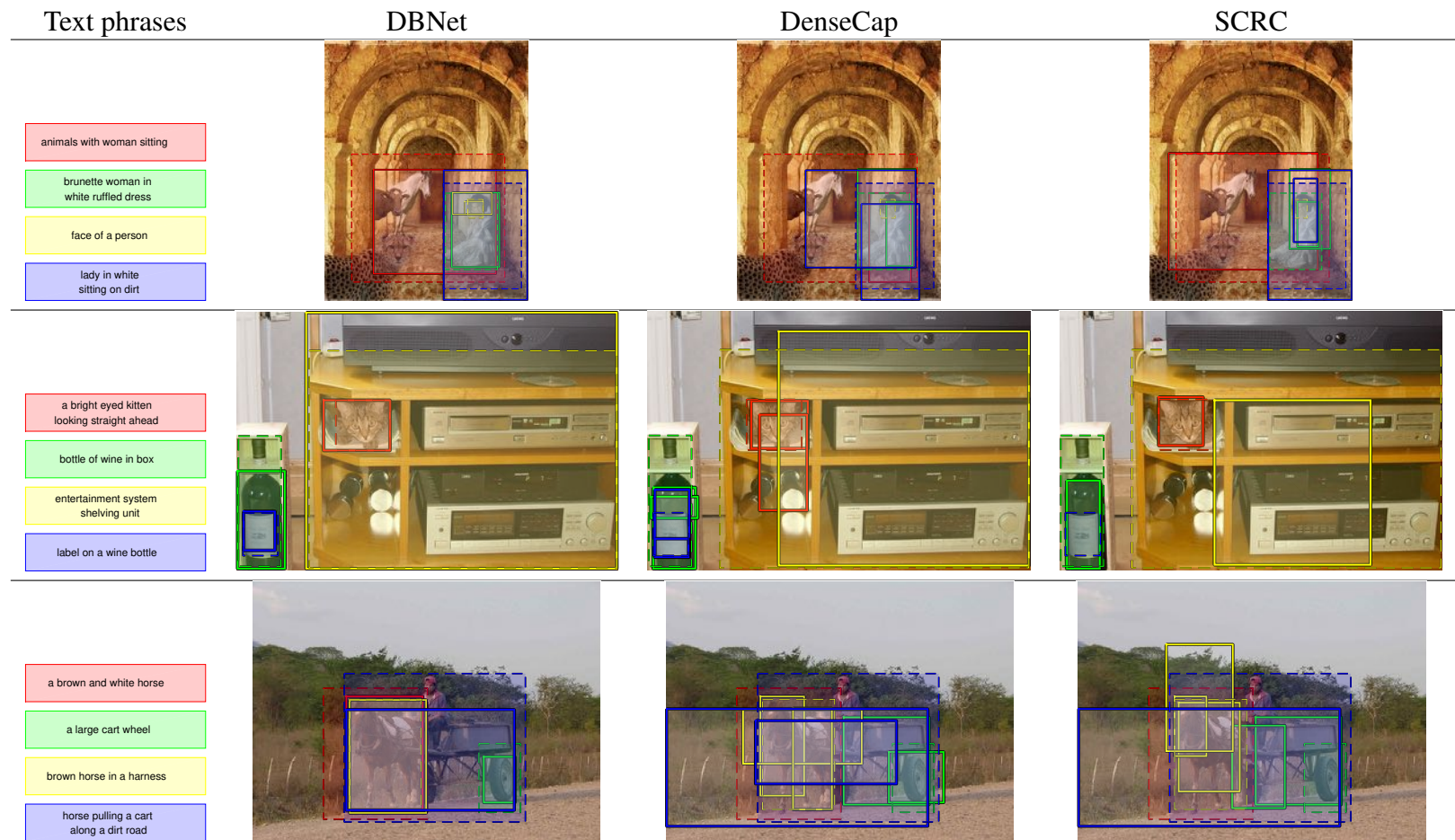
Figure 14: (continued from Figure 13) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. 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
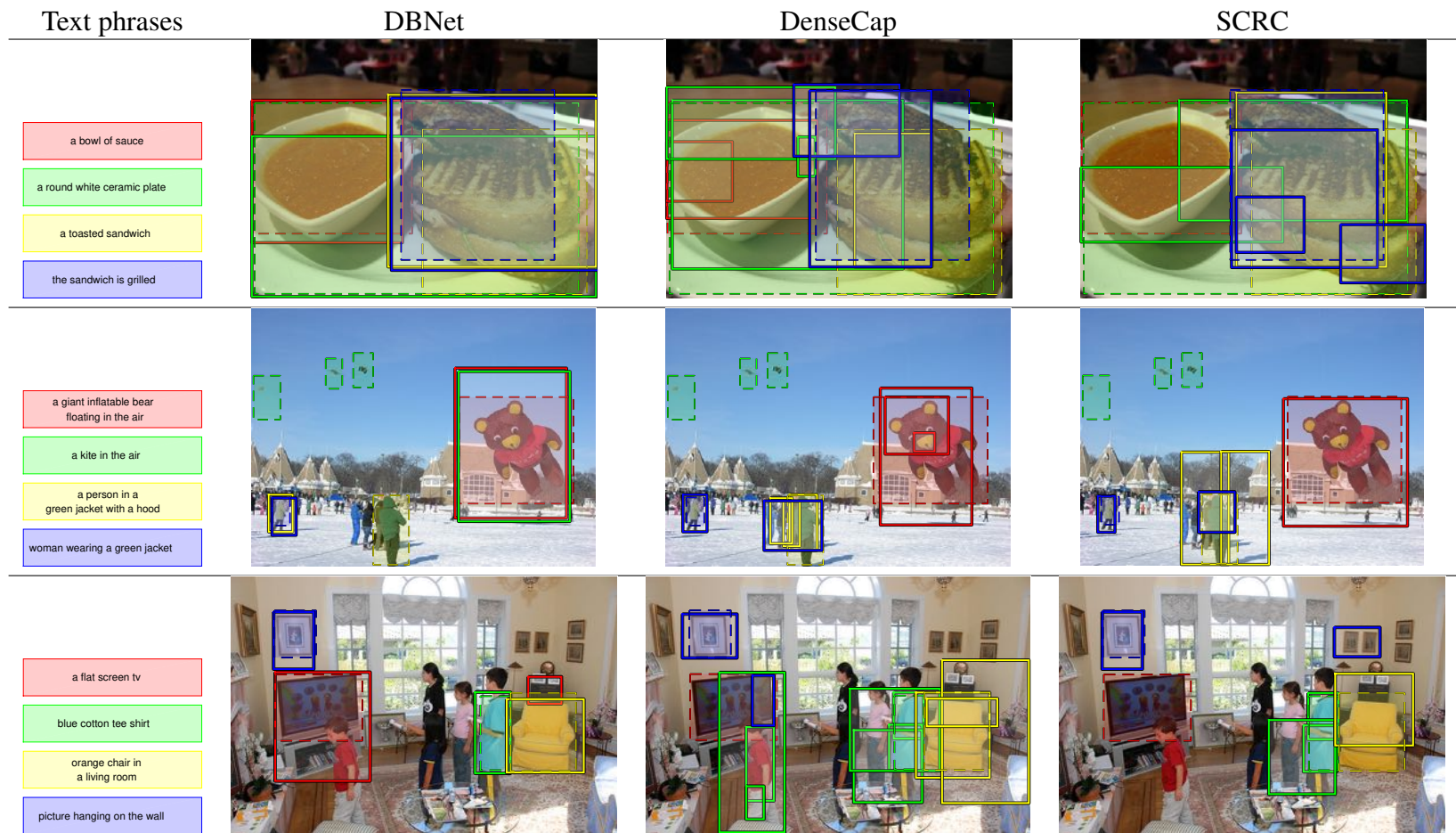
Figure 15: (continued from Figure 14) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.
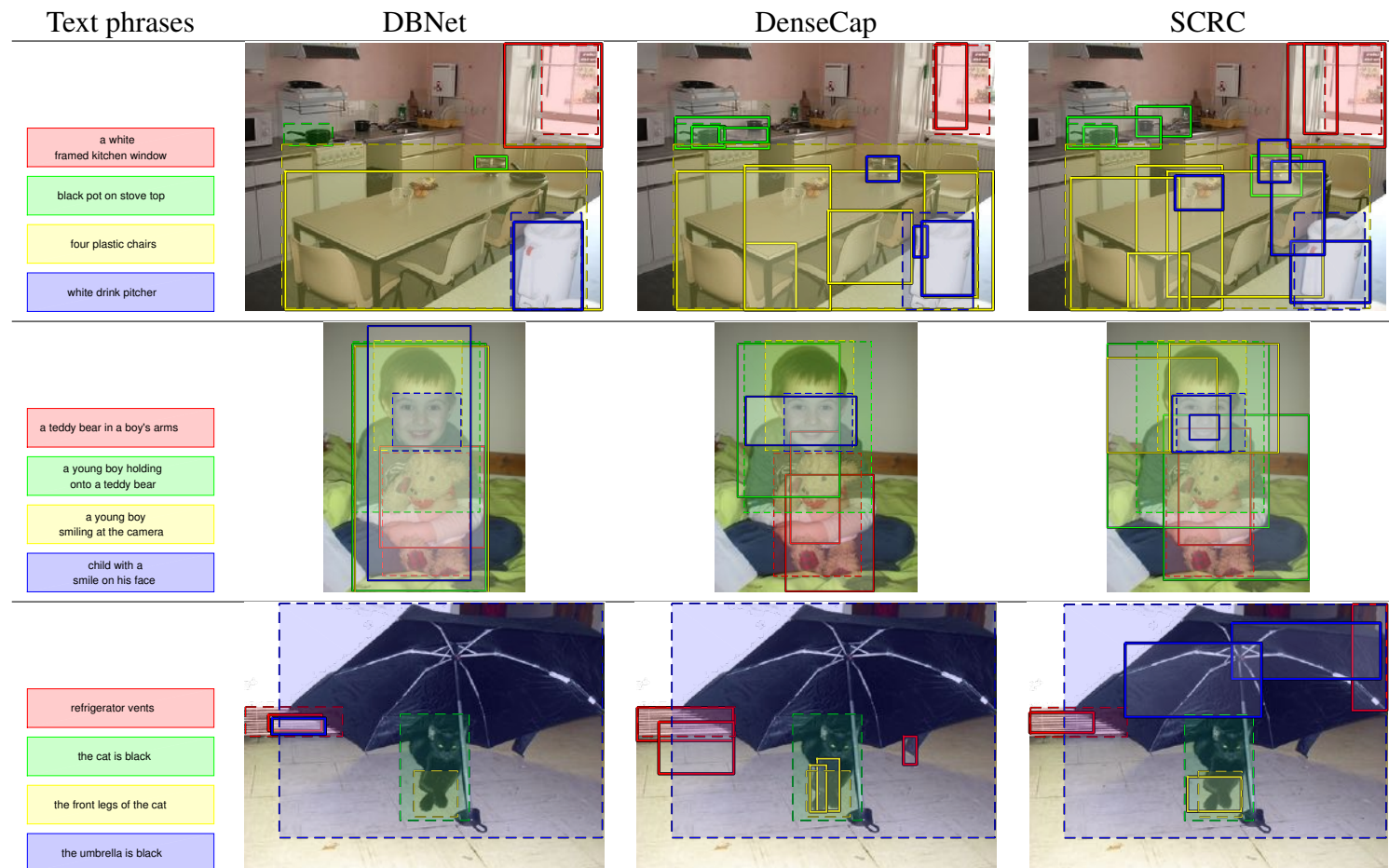
Figure 16: (continued from Figure 15) Qualitative detection results of DBNet, DenseCap, and SCRC using phrase-dependent detection threshold. Detection results of four different text phrases are shown for each image. The colors of the bounding boxes correspond to the colors of text phrases on the left. The semi-transparent boxes with dashed boundaries are ground truth regions, and the boxes with solid boundaries are detection results of three models.

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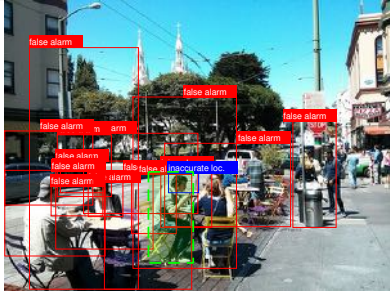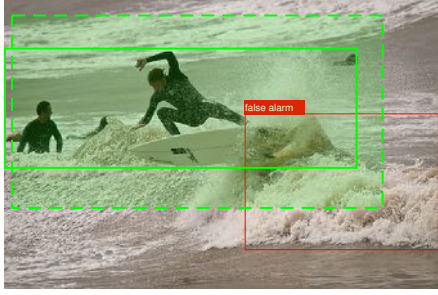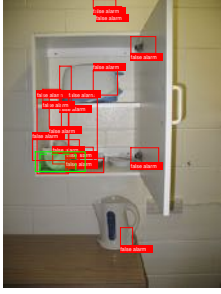
Text phrases	DBNet	DenseCap	SCRC
a man with dark hair eating outside			
a group of swimmers in the ocean			
a multi colored towel in the cabinet			

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 ($0 < IoU < 0.5$).
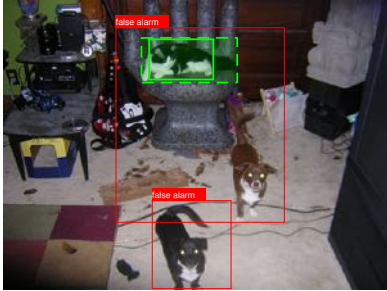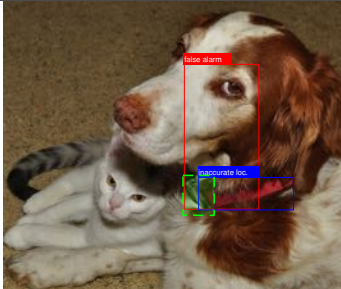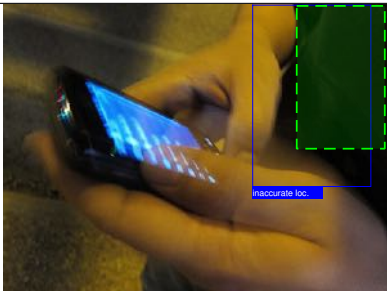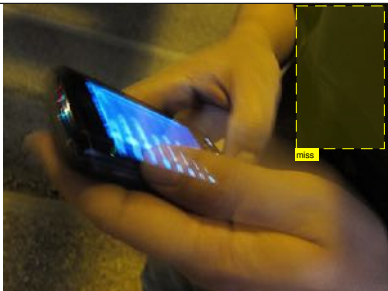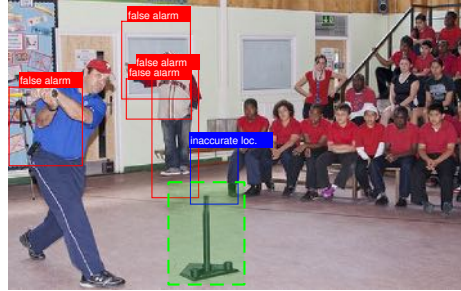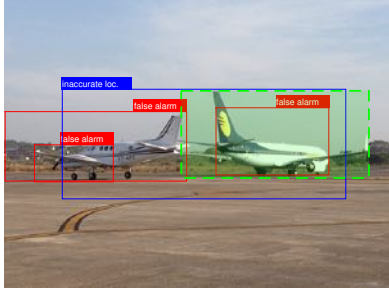
Text phrases	DBNet	DenseCap	SCRC
a black and white cat			
a buckle is on the collar			
a black shirt			

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 ($0 < IoU < 0.5$).
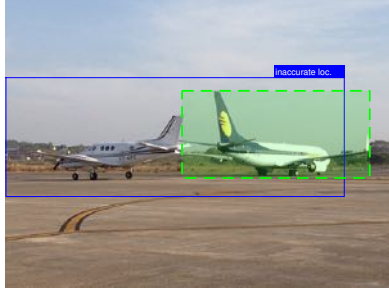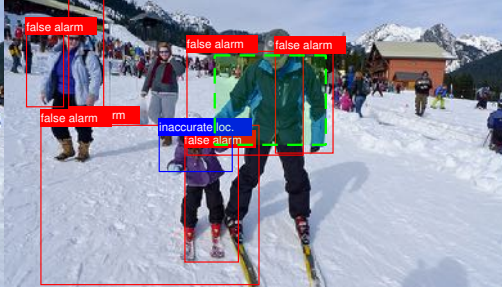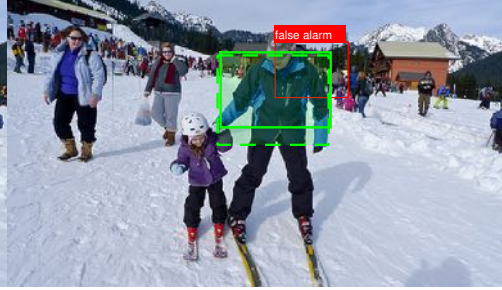
Text phrases	DBNet	DenseCap	SCRC
a baseball tee			
airplane parked on tarmac			
a 2 toned blue winter jacket			

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.
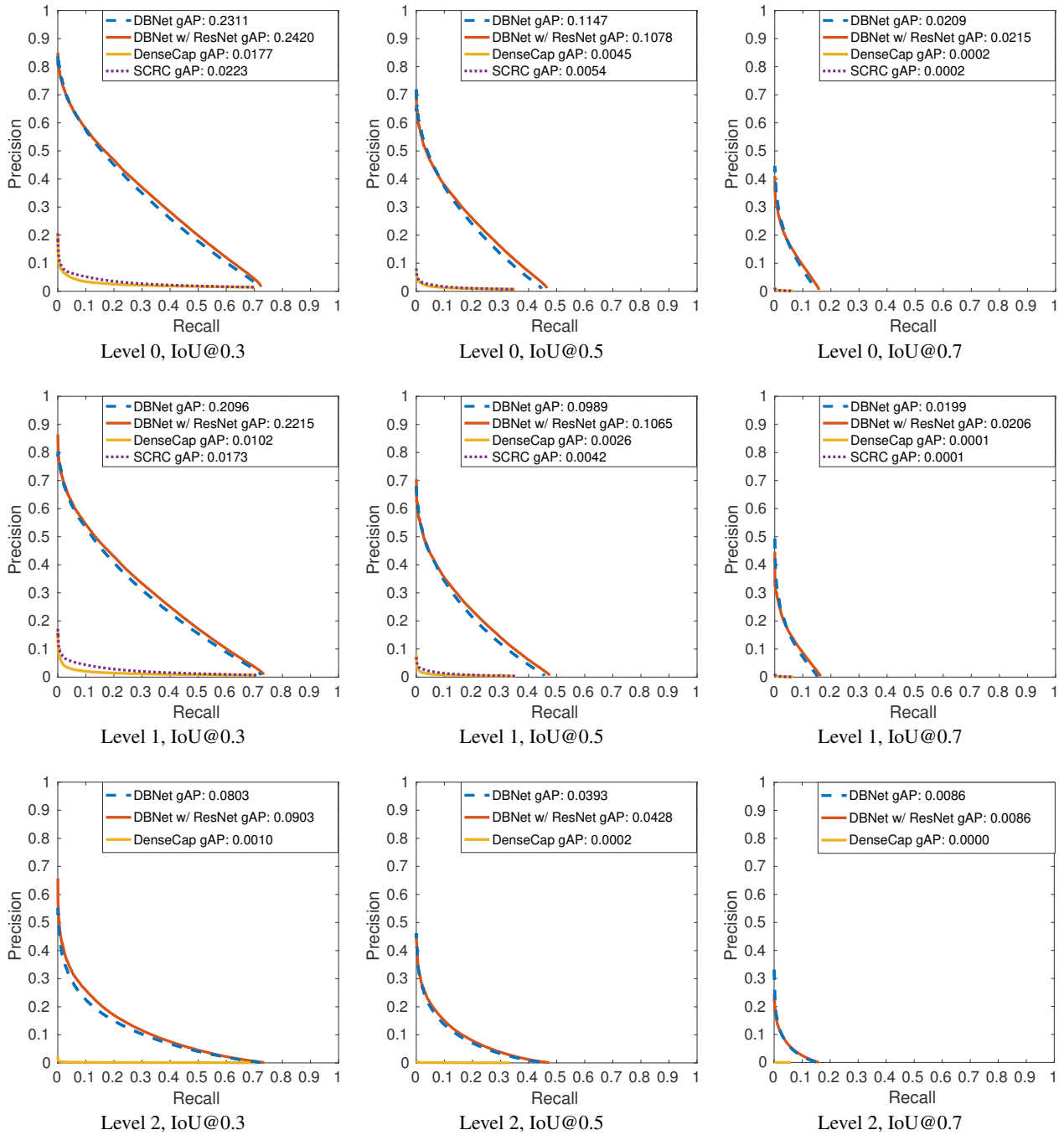

Figure 23: Phrase-independent precision-recall curves for calculating gAP.

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.
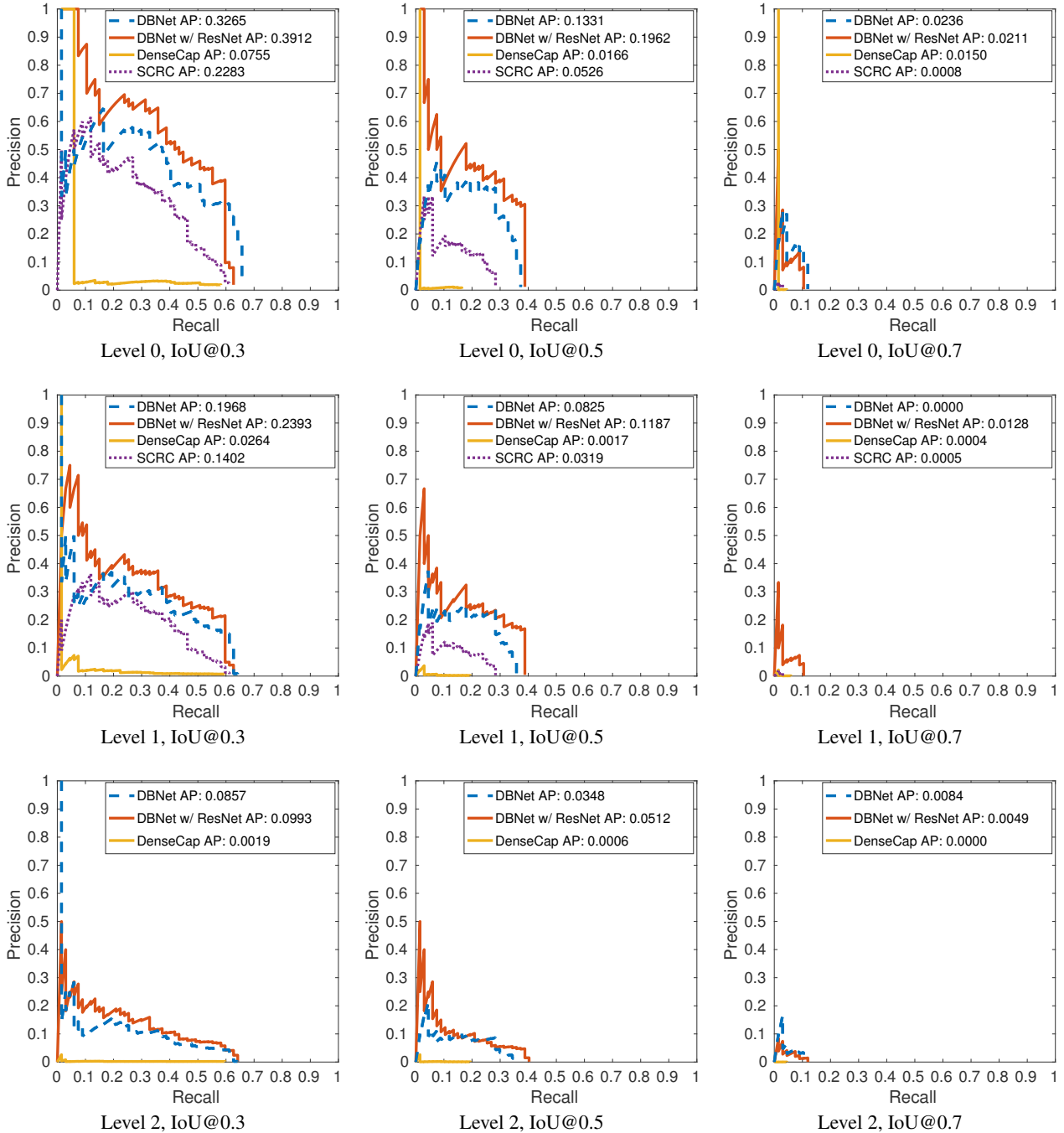

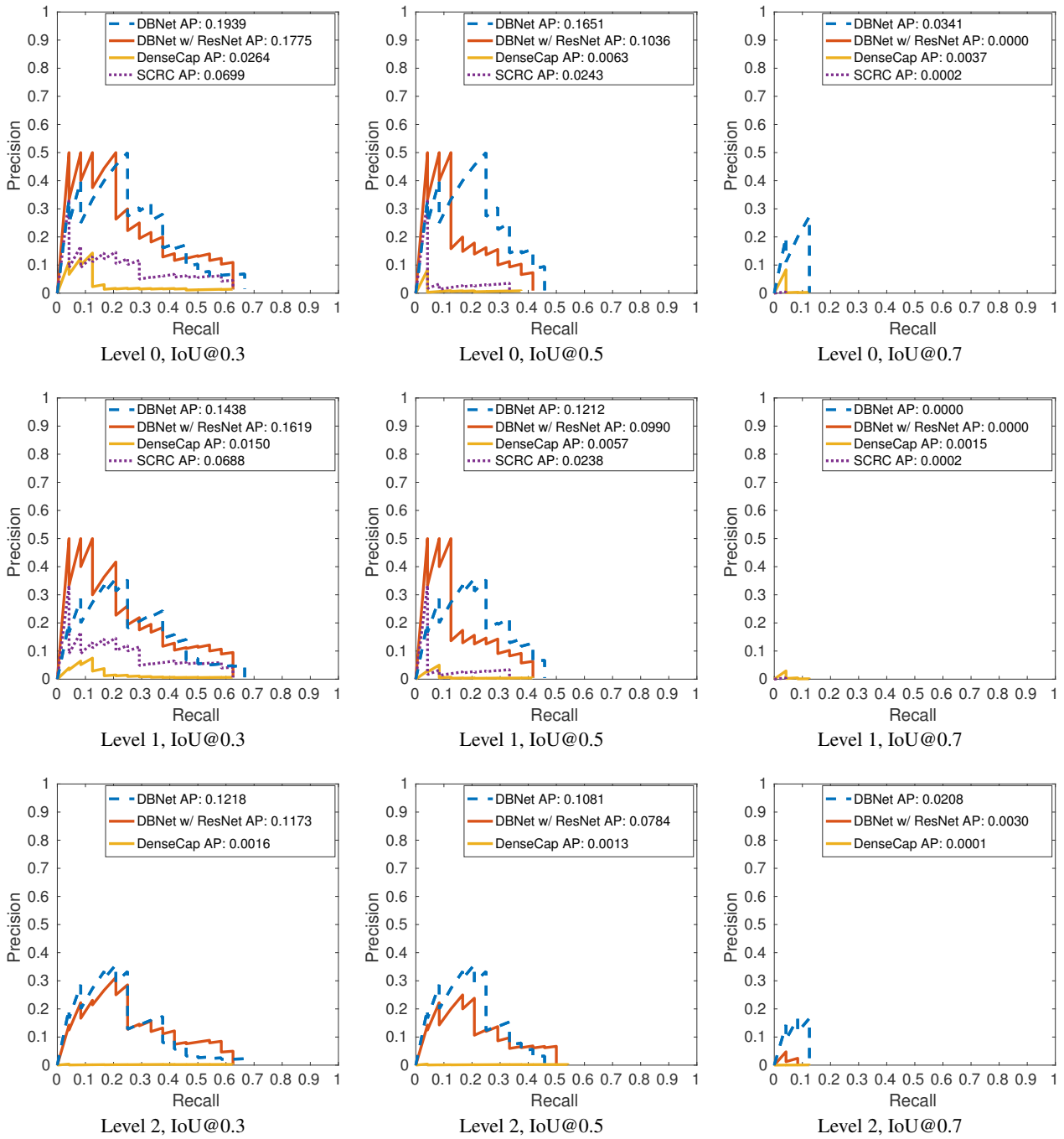**Figure 24:** Precision-recall curves for text phrase "head of a person".

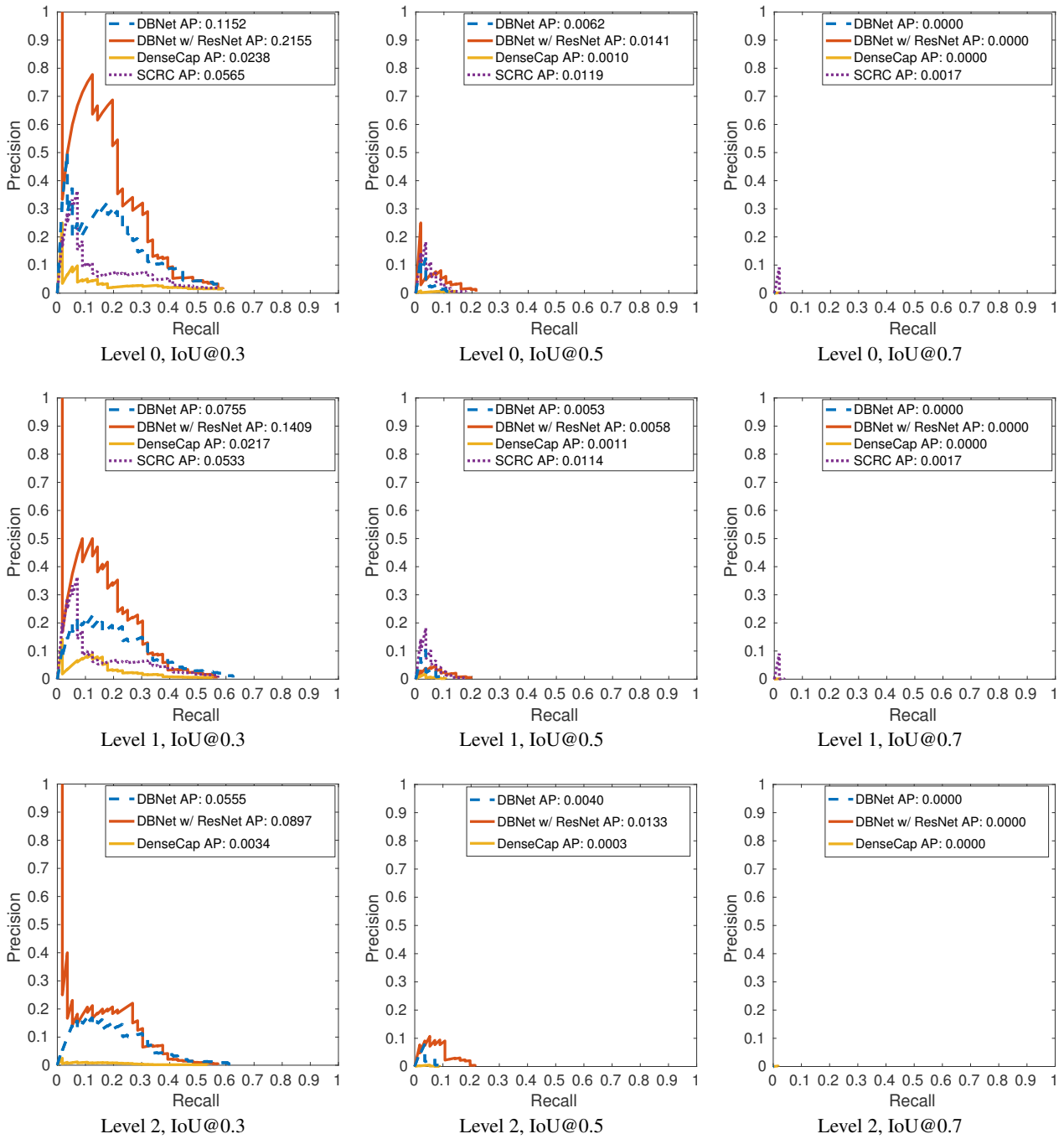**Figure 25:** Precision-recall curves for text phrase "a window on the building".

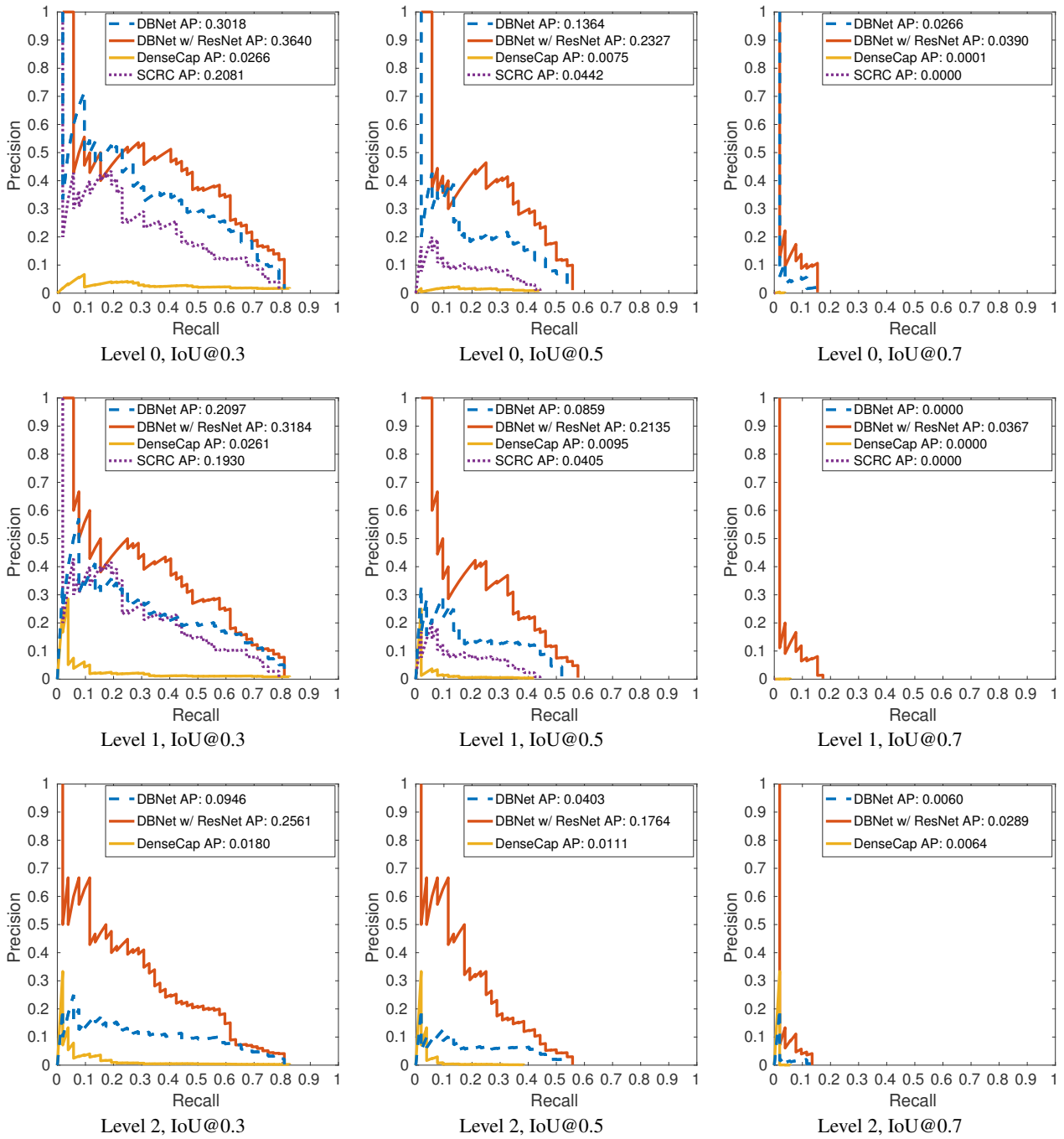**Figure 26:** Precision-recall curves for text phrase "the water is calm".

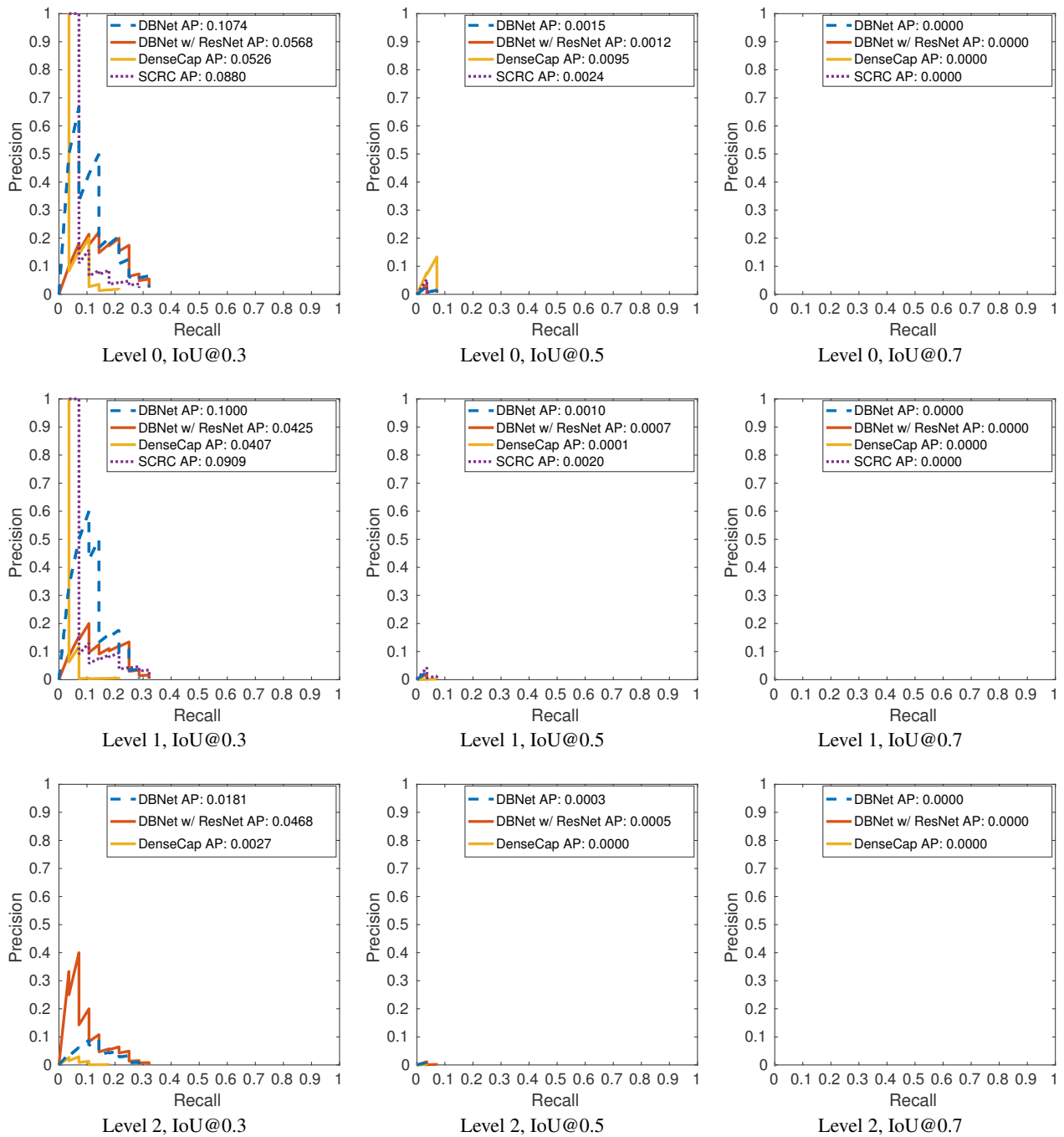**Figure 27:** Precision-recall curves for text phrase "man wearing blue jeans".

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