Object Localization with Attribute Preference based on Top-Down Attention

Soubarna Banik[1][0000−0002−2990−608X], Mikko Lauri[2][0000−0002−2223−9253], Alois Knoll[1][0000−0003−4840−076X], and Simone Frintrop[2][0000−0002−9475−3593]

[1] Technical University of Munich, Munich, Germany
soubarna.banik@tum.de,knoll@in.tum.de
[2] University of Hamburg, Hamburg, Germany
{lauri, frintrop}@informatik.uni-hamburg.de

Abstract. We propose a weakly-supervised approach for object localization based on top-down attention which is able to consider both attributes and object classes as attentional cues. This enables to not only search for objects but additionally for objects with specific attributes such as colors or shapes. Our approach consists of two streams: an attribute stream and an object stream. By tracing backward through these two streams and localizing activated neurons in hidden layers, we generate two top-down attention maps, one for attributes and one for objects. Fusing these maps generates a joint attention map, which highlights regions with a specific attribute and object. We show experimentally that our method can localize objects in cluttered images by only specifying their attributes, and that instances of the same class can be discriminated based on their attributes.

Keywords: Object localization · Object attribute · Top-down attention.

1 Introduction

Unconstrained environments, unknown objects, varying illumination, and limited computational resources challenge vision algorithms in real-world applications such as robotics. In human perception, attention mechanisms enable to cope with such challenges by prioritizing processing on the most relevant information [21]. For humans, the most relevant parts of a scene are determined by bottom-up and top-down factors [28]. Bottom-up factors guide our attention to salient parts of a scene that automatically draw attention, such as a flickering light or a strong color contrast. Top-down attention enables us to focus on behaviorally relevant regions: the bakery when we are hungry, or the station clock when hurrying to catch our train.

While many computational models of bottom-up attention (saliency) have been proposed [9, 13, 14], top-down attention is not as well investigated. Traditional top-down attention systems have mainly emphasized and localized target-specific features such as a certain color or orientation [8, 19]. Recent top-down attention approaches based on convolutional neural networks (CNNs) localize

(a) Input (b) Object (c) Object + Attribute

Fig. 1. (a) Input image, and top-down attention maps for (b) an object (dog), and for (c) a combination of both object and attribute (brown dog).

objects by tracing back activations from class-specific neurons to hidden layer neurons [3, 39]. This enables to quickly find the panda [3] or any other pre-trained object in a scene.

Existing methods search for either features [8, 19] or objects [3, 39], but it is known that both cues are important targets in human visual attention [17, 37][3]: humans may look for a specific object class, but also guide their attention to attributes, for example to all red things or to striped objects. In applications such as service robotics or human-robot interaction, a combination of both is important: a robot should be able to not only focus on all cups on a table, but to select a cup with certain attributes ("Bring me the blue cup"). Additionally, a task a human gives to a robot does not necessarily include the object class, but sometimes only an attribute ("Pick up the red item over there").

In this paper, we propose a weakly supervised method for object localization using a top-down attention mechanism which is able to consider both attributes and object class information. The input to our system is an image and a desired target attribute and object, e.g., "brown" and "dog" in Fig. 1. The outputs of the system are an attribute- and an object-specific attention map. The attribute and object attention maps can be used separately, or they can be fused to obtain a combined attribute-object attention map that highlights the regions in the input image that correspond to both the desired attribute and object. Our approach is able to localize attributes in images, as well as discriminate between object instances. In attribute localization, we show our method performs better empirically than general visual question answering [35].

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes the architecture of our proposed approach and Section 4 presents the experimental results. Concluding remarks are given in Section 5.

2 Related work

We briefly review relevant works in attribute classification and localization, top-down attention, weakly supervised object localization, and image captioning

[3] A third cue is spatial location, which we do not address here.

and visual question answering. We focus on how top-down attention is treated in these works, motivating our proposed approach.

Attribute classification and localization: Attributes can describe parts of an object (*has nose*, *has headlight*) [6], or the characteristic features of objects (*color*, *texture*, *pattern*) [2, 12, 24]. Existing attribute classification methods such as [5, 18, 22] use CNN extracted feature descriptors and support vector machines. In addition, approaches applying Bayesian theory [7, 15] and deep learning [2, 4, 33, 40] have been proposed. For attribute localization, Xiao et al. [34] focus on relative attributes between pairs of images to learn the spatial extent of the attributes. Wang et al. [30] propose a weakly supervised method for studying scene configuration and simultaneous attribute localization. However, these methods do not support adding a top-down signal or querying for an attribute.

Top-down attention: Different studies find that humans can direct their attention to not only objects, but also to spatial locations and features [17, 37]. The guided search model of Wolfe [31] introduces top-down cues that boost target-relevant features. Inspired by these psychological theories, early computational models of top-down attention learn how much feature channels contribute to finding a target object. Some of these approaches excite target related features [39] while others also inhibit irrelevant features [8, 29].

Deep-learning-based top-down object search [3, 39] is realized by highlighting regions corresponding to the target class in the hidden layer activation maps. Cao et al. [3] introduced feedback layers in a classification network to infer the activation status of hidden layer neurons. During the backward pass, neurons in the feedback layers act as gates and open only connections associated with the target class. Zhang et al. [39] proposed Excitation Backprop (Ex-BP), which given a top-down signal for a target object, reveals the location of the target-specific activation in the input image. The algorithm applies to any deep classification network without the need for any modification, unlike [3].

Weakly supervised object localization: Segmentation-based approaches for weakly supervised localization [10, 30] classify each previously determined image segment, thereby giving the location information along with the object class. Oquab's fully convolutional model [20] is similar to the segmentation-based approaches, except instead of segments each sliding window is classified. However, the aforementioned approaches do not facilitate adding a top-down signal for a target object. Alternatively, recovering the location information from classification CNNs allows a mechanism for incorporating the top-down signal [25, 41]. Approaches like Class Activation Mapping (CAM) [41] and Grad-CAM [25] are similar to Ex-BP [39] in functionality.

Image captioning and visual question answering: Some works in image captioning (IC) and visual question answering (VQA) learn to output attention maps corresponding to specific words in the generated caption or answer sentence [1, 16, 26, 36]. The generated sentences may include object names, attributes, or relationships and the corresponding attention map shows the relevant region in the input image. However, these methods are not suited for generic object or attribute attention. Top-down signals in IC [16, 36] are limited to the

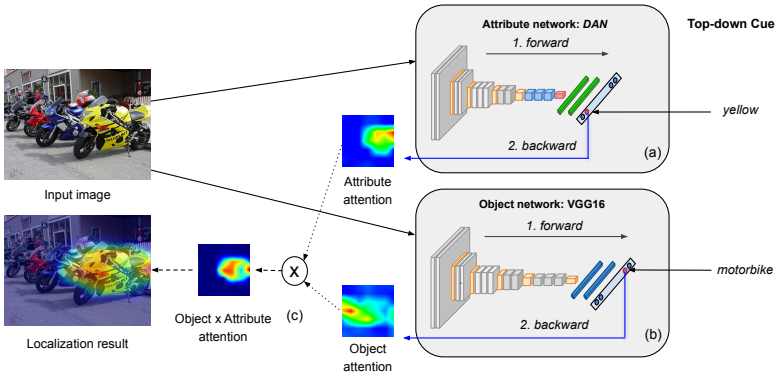

Fig. 2. Overview of the object localization process using attributes. Attention maps are generated from (a) the attribute network for the target attribute *(yellow)*, and from (b) the object network for the target object *(motorbike)*. (c) Object and attribute attention maps are combined to obtain the object-attribute attention map, showing the location of the target object having the target attribute (yellow motorbike).

set of words from the generated caption. VQA methods such as [1, 26] rely on question-answer annotations [11] and an object detection network that uses object and attribute annotations and their corresponding locations.

3 Architecture

Fig. 2 shows an overview of our object localization process with attribute guidance. Our system contains two streams: an attribute classification stream ((a) in Fig. 2) trained to classify attributes such as color, shape, pattern, and texture; and an object classification stream (b) consisting of a standard object classification network, namely the VGG16 network [27]. The attribute network, named Deep Attribute Network (DAN) is described in Sec. 3.1. In both streams, a forward pass enables classification of attributes and objects, and a backward pass traces the activations in a top-down manner to output an attribute- and an object-specific attention map, as described in Sec. 3.2. The attribute and object attention maps can be used separately, or they can be combined to obtain an attribute-object attention map (c), which shows the locations in the input image that correspond to both the desired attribute and object.

3.1 Deep attribute network

Attributes are descriptive properties of objects [6]. Low-level attributes such as color, shape, or type of material can be shared by objects of different classes. High-level object-part attributes, such as headlight, wheel, side mirror, are specific to certain objects and are not universally applicable. We consider only low-level attributes that can describe a wide range of objects. CNNs learn Gabor-like

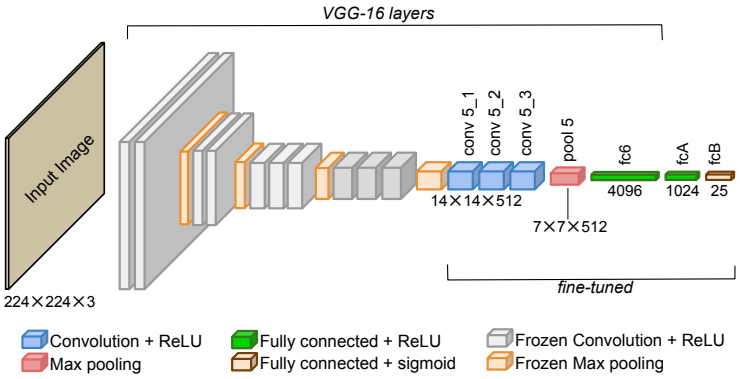

Fig. 3. The deep attribute network (DAN). The network is adapted from the VGG-16 model [27], by replacing the last two fully-connected layers with the new layers *fcA* and *fcB*. DAN is finetuned till layer *conv5_1*.

and color features in the first layers [38], which are useful for discriminating our target attribute groups - *color*, *shape*, *pattern* and *texture*. So, we modify VGG-16 [27], a standard object classification network, and finetune the last layers to adapt to the attribute classification similarly as in [2]. We obtain the Deep Attribute Network (DAN) for low-level attribute classification as shown in Fig. 3.

DAN is trained using transfer learning pre-trained on the ImageNet dataset [23]. All layers up to and including (*fc6*) are retrained. Two new fully-connected layers *fcA* with 1024 units and *fcB* with 25 units are added. The 25 units in the *fcB* layer correspond to the 25 target attribute classes in the ImageNet-Attribute dataset [24] in the attribute groups *color*, *shape*, *pattern* and *texture*. As object attribute classification is a multilabel classification task, the final layer *fcB* uses a sigmoid activation. All other layers use ReLU activation. We train DAN similarly as in [2], using weighted cross-entropy loss for attribute classification without any localization labels.

3.2 Top-down attention

Top-down attention localizes image regions likely to show the target object. In deep networks, target information is usually provided to the network at the output layer by indicating the class nodes which are target relevant, e.g. [3, 41]. We use Grad-CAM [25], originally introduced to provide visual explanations for deep networks, to generate top-down attention maps. Given a top-down cue at an output node of the network, Grad-CAM generates an attention map highlighting the regions in the input image that activated the output node.

Grad-CAM calculates the top-down attention map for a target class c as a weighted sum of activation maps A_k in the last convolutional layer. The weights α_k^c are derived from the gradients of the output element corresponding to c with respect to elements of the activation map A_k. The weight indicates the

importance of the activation map for the target class. The attention map L^c for class c is $L^c = \text{ReLU}\left(\sum_k \alpha_k^c A_k\right)$, where the ReLU ensures only positive contributions towards the target class c are considered.

We compute the attribute and object specific attention maps using Grad-CAM. To identify a target object with a specific attribute, the attribute-object attention map is computed by multiplying the individual maps as shown in Fig. 2 (c). The greatest value in an attention map indicates the object location. As the attribute and object streams are independent, our method also works for unknown object classes, in which case only the attribute cue is used.

4 Experiments

We introduce the datasets we use in Sec. 4.1 and describe our evaluation methodology in Sec. 4.2. Sec. 4.3 demonstrates that object attributes can be successfully applied to localize objects in images. Sec. 4.4 then shows how much attributes help in object localization. We omit evaluation of object localization without attributes as this is covered in previous works [27].

4.1 Datasets

ImageNet-Attribute [24]: Contains annotations of 25 low-level object attributes in four categories: color , shape , pattern , and texture. The dataset consists of 9600 images from 384 synsets. Each image contains one object and is also annotated with the object class label and the corresponding bounding box. We divide the dataset into training, validation, and test sets with 5760, 1920, and 1920 images, respectively. DAN is trained on the training split of this dataset. The dataset is useful for training attributes, but not especially suitable for evaluating object or attribute localization since the images show mostly only a single object in large scale, often centered.

a-Pascal [6]: The images contain a varying number of objects from 20 classes annotated with bounding boxes, and 64 attributes describing shape, material, and high-level object components. The dataset is well suited for the attribute-object localization task. For our evaluation, we select the low-level attributes that are present in the training dataset: three shape attributes (*2D boxy, 3D boxy, round*) and five material attributes (*metal, wood, furry, shiny, vegetation*). The shape attributes *2D boxy* and *3D boxy* are not present in the training dataset. After manually reviewing the images, we found both attributes similar in appearance to the *rectangle* attribute of ImageNet-Attribute dataset. We merge these two attributes and evaluate them as the *rectangle* attribute.

Object-Attribute: We create a dataset for evaluating the attribute localization task. The new Object-Attribute dataset consists of 60 images from a cluttered kitchen environment. The number of objects per image is between 3 and 8. We annotate the images using the same color and shape attributes as in the ImageNet-Attribute dataset, and add bounding box annotations for objects.

4.2 Pointing game evaluation

Top-down attention maps highlight class-discriminative regions but do not provide pixel-precise boundaries of target classes. Due to this, we evaluate the performance of our approach with the *Pointing Game* experiment [39]. To calculate accuracy, the maximum saliency point (MSP) in a target-specific top-down attention map is selected. A hit is counted if the point is on any of the annotated instances of the cued target, otherwise a miss occurs. The pointing game accuracy is calculated as $hits/(hits + misses)$. This is equivalent to the top-1 accuracy metric used for evaluating object localization methods [27]. We also calculate the accuracy in a more relaxed setting from the top-3 most salient regions. The attention map is thresholded at 90% of the maximum saliency value of the map. The three regions with the greatest average saliency are selected. If any of the centroids of these regions lie in the ground truth bounding box area, a hit is counted. Otherwise, a miss is counted.

To calculate recall, we determine a hit or a miss similarly as for top-1 accuracy. We inhibit the area around the current MSP, and extract the next MSP, repeating k times. For top-k recall, we count the number of unique bounding boxes with hits among the first k MSPs, and calculate the ratio to the total number of applicable bounding boxes. We average across all images.

4.3 Experiment 1: Attribute localization

Can attributes be used to localize objects without any information about the class of the target object? This is of interest for autonomous robots to find objects only based on their attributes such as with the given task "bring me the red object over there". We first compare variants of our attribute localization method using either Grad-CAM [25], Excitation-Backprop (Ex-BP) [39], or Class Activation Mapping (CAM) [41]. We also compare the attribute localization performance to a popular VQA method.

Pointing game accuracy and recall: We generate attention maps at *pool5* of DAN, for Grad-CAM and Ex-BP. For CAM, we follow [41]: first, layers *(pool5-fcB)* in DAN are removed and a convolutional layer *conv6* with 512 filters of size 3×3, stride 1, padding 1 is added. We insert a global average pooling layer, a fully-connected layer *fcC* with 25 nodes and a sigmoid layer. The newly added layers *(conv6-fcC)* are trained for 6 epochs similarly as in [2]. We generate the attention maps for CAM at layer *conv6* of the modified network.

Table 1 reports the accuracy and recall of Ex-BP, CAM, and Grad-CAM on all datasets. Only attribute classes are used as top-down cues and the mean across all attribute classes is reported[4]. For all three methods, recall on the Imagenet Attribute dataset is high, as is the top-1 accuracy. Among the test datasets, CAM and Grad-CAM achieve similar accuracy and recall on Object-Attribute dataset, whereas, in the more complex a-Pascal dataset, Grad-CAM performs the best on our model. Overall the result indicates that objects can be

[4] The types and number of attributes are different for the three datasets; see Sect. 4.1.

Table 1. Pointing game: mean attribute localization accuracy and recall for Imagenet-Attribute (IA), Object-Attribute (OA) and a-Pascal datasets.

	Accuracy						Recall					
	IA		OA		a-Pascal		IA		OA		a-Pascal	
	top-1	top-3	top-1	top-3	top-1	top-3	top-1	top-10	top-1	top-10	top-1	top-10
Ex-BP	0.88	0.90	0.65	0.72	0.52	0.57	0.88	0.88	0.58	0.61	0.44	0.45
CAM	0.85	0.88	0.69	0.75	0.47	0.52	0.87	0.87	0.62	0.64	0.42	0.43
Grad-CAM	0.79	0.83	0.70	0.71	0.56	0.60	0.80	0.80	0.61	0.64	0.48	0.50

Fig. 4. Sample attribute attention maps overlapped on the input image from (a) a-Pascal and (b) Object-Attribute datasets.

localized using only their attributes. Based on the test dataset performance, we conclude that Grad-CAM performs best for generating attention maps. Grad-CAM is exclusively used in all of the remaining experiments in the paper.

Fig. 4 shows sample attention maps for the attributes *round*, *vegetation*, *red* and *green* from the a-Pascal and Object-Attribute datasets. The bounding box of the object with the target attribute is marked in red. The maximum saliency point, indicated by the red dot, shows the predicted location of the objects, even if they are partially occluded (the bucket in the rightmost image in Fig. 4 (b)).

Comparison to Visual Question Answering: Since visual question answering methods can also be used for producing attention map for a specific query, we compare to one such method named "Ask, Attend and Answer" (AAA) [35]. We input an image and a query "Where is the x object?", where x is the attribute label. Table 2 reports the accuracy and recall on Object-Attribute dataset for all attributes, and separately for the attribute groups color and shape. This dataset is not used to finetune any of the models. Our simple method that focuses on attribute localization clearly outperforms the more general VQA method. AAA performs better on the shape attributes, particularly for the shapes 'long' and 'rectangle'. In contrast to AAA, our method is weakly supervised for object localization, using object and attribute annotation, but no location annotation.

4.4 Experiment 2: Joint attribute-object localization

Does using attributes help to find objects? To answer this, we conduct the pointing game experiment for a combined object and attribute top-down signal, using

Table 2. Comparison of attribute localization performance between our approach and Visual Question Answering method (AAA) on the Object-Attribute dataset.

	top-1 accuracy			top-1 recall			top-10 recall		
attribute type	all	color	shape	all	color	shape	all	color	shape
AAA [35]	0.55	0.53	0.59	0.48	0.48	0.47	0.50	0.48	0.55
Ours	**0.70**	0.75	0.55	**0.61**	0.69	0.38	**0.64**	0.72	0.41

Table 3. Pointing game: top-1 attribute-object localization accuracy. Number in parentheses shows improvement compared to object localization. Attribute attention improves accuracy for a majority of object classes, as highlighted in blue.

	loc acc →	0.84	0.65	0.58	0.48	0.41	0.41	0.35	
Obj/Attrib	loc acc ↓	furry	round	metallic	wooden	shiny	vegetation	rectangular	avg. improvement
cat	0.96	0.94 (0)	-	-	-	-	-	-	0.00
dog	0.90	0.90 (0.01)	-	-	-	-	-	-	0.01
cow	0.90	0.95 (0.06)	-	-	-	-	-	-	0.06
horse	0.86	0.93 (0.06)	-	-	-	-	-	-	0.06
motorbike	0.83	-	-	0.74 (-0.06)	-	0.67 (-0.03)	-	-	-0.05
aeroplane	0.82	-	1.00 (0)	0.89 (0.08)	0.60 (0)	0.83 (0.02)	-	0.68 (-0.04)	0.01
train	0.82	-	0.78 (0.11)	0.87 (0.03)	-	0.76 (-0.06)	-	0.81 (-0.03)	0.01
bus	0.81	-	-	0.78 (-0.04)	-	0.59 (-0.11)	-	0.58 (-0.15)	-0.10
diningtable	0.78	-	0.75 (-0.06)	0.80 (0.10)	0.68 (-0.06)	0.81 (0)	-	0.77 (-0.06)	-0.02
sofa	0.76	-	0.55 (-0.18)	-	0.69 (-0.08)	-	-	0.52 (-0.19)	-0.15
tvmonitor	0.75	-	-	0.48 (-0.19)	-	0.52 (-0.23)	-	0.54 (-0.20)	-0.21
bicycle	0.71	-	-	0.76 (0.04)	-	0.66 (0)	-	-	0.02
boat	0.59	-	0.55 (0)	0.66 (0.17)	0.50 (0.06)	0.45 (0.09)	-	0.57 (0.03)	0.07
car	0.57	-	-	0.60 (0.06)	-	0.47 (-0.04)	-	0.47 (-0.03)	0.00
pottedplant	0.55	-	-	-	-	-	0.56 (0.01)	-	0.01
chair	0.48	-	-	0.31 (0.06)	0.52 (0.03)	0.48 (0.09)	-	0.44 (0.04)	0.06
bottle	0.41	-	-	0.44 (0.11)	-	0.35 (0.02)	-	-	0.07

Grad-CAM to create the attention maps. We evaluate on the a-Pascal dataset or its subsets. For object localization, we replace the softmax layer of VGG-16 [27] with a sigmoid layer with 20 output nodes corresponding to the object classes, and train with the cross-entropy loss.

Object vs attribute-object localization: Table 3 reports the accuracy for the integrated attribute-object attention and the improvement over only using object attention in brackets. The average improvement is reported in the rightmost column. The object/attribute cases where the localization improved compared to only using object attention are highlighted in blue. The additional attribute cue improves the localization for 10 out of 17 classes. Accuracy decreases for 5 classes - *motorbike*, *bus*, *diningtable*, *sofa* and *tvmonitor*. This may be due to a dominance of the object classification stream which has been trained with a much larger dataset. For example, the greatest decreases are observed for the attributes *shiny* and *rectangle*. The average precision of DAN on a-Pascal dataset for these two attributes are 0.32 and 0.55 respectively [2]. The poor attribute classification performance affects the corresponding attribute localization accuracy (0.41 and 0.35 respectively) and therefore, the combined object and attribute localization performance. Other failure cases are for the object/attribute combinations *sofa/round* and *tvmonitor/metallic*. This is explained by the contradictory annotations in the training (*rectangle*) and test dataset (*round*) for

Table 4. Localization performance in images with visually similar objects. Attribute attention clearly improves performance.

Top-down cue	top-1 accuracy	top-1 recall	top-10 recall
Object	0.56	0.33	0.35
Attribute-Object	0.64	0.39	0.41

Table 5. Attribute-object localization performance for combinations of objects and attributes either in the training set or not.

Combination type	top-1 accuracy	top-1 recall	top-10 recall
In training set	0.61	0.52	0.54
Not in training set	0.58	0.49	0.51

sofa, and by the absence of the object-attribute combination *tvmonitor-metallic* in the training dataset respectively. The overall performance in this experiment clearly shows that attributes are useful cues for localizing objects within the proposed top-down attention framework.

Searching amongst visually similar objects: We consider images with multiple instances of the target attribute with a different object class, e.g., images with both a red car and a red ball. These search cases are the most difficult: in psychological visual search tasks, humans are slower to find the target if there are distractors with similar appearance to the target [32]. We repeat the previous experiment on 287 such images in a-Pascal. Table 4 reports the accuracy and recall for object and attribute-object attention. Corresponding to the more difficult task, attribute-object localization performance is naturally lower in these cases than the average of the dataset (0.67). Nevertheless, localization accuracy and recall improve when an attribute top-down signal is used in addition to the class information. This shows a clear advantage of the attribute information for object localization.

Unobserved object-attribute combinations: We find the combinations of objects and attributes that exist both in the training and test data, and also the combinations that appear only in the test data but not in the training data. We evaluate attribute-object localization performance on both of these subsets. Table 5 shows that the performance for object-attribute pairs that are not present in the training data decreases, as expected, but only by 3% compared to the cases present in the training data. This shows that our method generalizes well to unseen pairs of objects and attributes.

5 Conclusion

We present a simple yet effective approach for localizing attributes in images with top-down attention. We generate attribute and object attention maps to localize attributes, objects, or a combination of both. Our approach can search

for objects, or for image regions of certain properties, and discriminate object class instances based on attributes, while generalizing to unseen combinations of objects and attributes. A limitation of our method is that the contribution of object and attribute localization streams is not controllable. This may lead to a localization failure when either one of the streams fails. Further experiments can be conducted to demonstrate applicability in a real-world robotics scenario. Finally, a comparison to object localization methods could further help to demonstrate the benefits and disadvantages of attribute-based localization.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
2. Banik, S., Lauri, M., Frintrop, S.: Multi-label object attribute classification using a convolutional neural network. arXiv preprint arXiv:1811.04309 (2018)
3. Cao, C., Liu, X., Yang, Y., Yu, Y., Wang, J., Wang, Z., Huang, Y., Wang, L., Huang, C., Xu, W., et al.: Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In: ICCV (2015)
4. Chung, J., Lee, D., Seo, Y., Yoo, C.D.: Deep Attribute Networks. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2012)
5. Cimpoi, M., Maji, S., Vedaldi, A.: Deep Filter Banks for Texture Recognition and Segmentation. In: CVPR (2015)
6. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing Objects by their Attributes. In: CVPR (2009)
7. Ferrari, V., Zisserman, A.: Learning Visual Attributes. In: NIPS (2008)
8. Frintrop, S., Backer, G., Rome, E.: Goal-directed search with a top-down modulated computational attention system. In: DAGM Symposium on Pattern Recognition (2005)
9. Frintrop, S., Werner, T., Martin Garcia, G.: Traditional saliency reloaded: A good old model in new shape. In: CVPR (2015)
10. Galleguillos, C., Babenko, B., Rabinovich, A., Belongie, S.: Weakly supervised object localization with stable segmentations. In: ECCV (2008)
11. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In: CVPR (2017)
12. Hu, C., Bai, X., Qi, L., Chen, P., Xue, G., Mei, L.: Vehicle Color Recognition with Spatial Pyramid Deep Learning. IEEE Trans. on Intell. Transp. Systems **16**(5), 2925–2934 (2015)
13. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. on PAMI **20**(11), 1254–1259 (1998)
14. Kümmerer, M., Wallis, T.S.A., Bethge, M.: Saliency benchmarking made easy: Separating models, maps and metrics. In: ECCV (2018)
15. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to Detect Unseen Object Classes by Between-class Attribute Transfer. In: CVPR (2009)
16. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: ECCV (2018)
17. Liu, T., Slotnick, S.D., Serences, J.T., Yantis, S.: Cortical mechanisms of feature-based attentional control. Cerebral Cortex **13**(12), 1334–1343 (2003)

18. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep Learning Face Attributes in the Wild. In: ICCV (2015)
19. Navalpakkam, V., Itti, L.: An integrated model of top-down and bottom-up attention for optimizing detection speed. In: CVPR (2006)
20. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: CVPR (2015)
21. Pashler, H.: The Psychology of Attention. MIT Press, Cambridge, MA (1997)
22. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: An astounding baseline for recognition. In: CVPR (2014)
23. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet Large scale Visual Recognition Challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
24. Russakovsky, O., Fei-Fei, L.: Attribute Learning in Large-scale Datasets. In: ECCV (2010)
25. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
26. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: CVPR (2016)
27. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-scale Image Recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Theeuwes, J.: Top-down and bottom-up control of visual selection. Acta Psychologica **135**, 77–99 (2010)
29. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. Artificial intelligence **78**(1-2), 507–545 (1995)
30. Wang, C., Ren, W., Huang, K., Tan, T.: Weakly supervised object localization with latent category learning. In: ECCV (2014)
31. Wolfe, J.M.: Guided search 2.0: A revised model of visual search. Psychonomic Bulletin and Review **1**(2), 202–238 (1994)
32. Wolfe, J.M.: Visual search. In: Attention. Psychology Press/Erlbaum (UK) Taylor & Francis (2016)
33. Wu, F., Wang, Z., Lu, W., Li, X., Yang, Y., Luo, J., Zhuang, Y.: Regularized Deep Belief Network for Image Attribute Detection. IEEE Transactions on Circuits and Systems for Video Technology **27**(7), 1464–1477 (2017)
34. Xiao, F., Lee, Y.J.: Localizing and visualizing relative attributes. In: Visual Attributes. Springer (2017)
35. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: ECCV. Springer (2016)
36. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)
37. Yantis, S., Serences, J.T.: Cortical mechanisms of space-based and object-based attentional control. Current Opinion in Neurobiology **13**(2), 187–193 (2003)
38. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in Deep Neural Networks? In: NIPS (2014)
39. Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down Neural Attention by Excitation Backprop. In: ECCV (2016)
40. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: Pose aligned Networks for Deep Attribute Modeling. In: CVPR (2014)
41. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)