Deep Attribute-preserving Metric Learning for Natural Language Object Retrieval

Jianan Li
Beijing Institute of Technology
20090964@bit.edu.cn

Yunchao Wei
National University of Singapore
eleweiyv@nus.edu.sg

Xiaodan Liang
Carnegie Mellon University
xiaodan1@cs.cmu.edu

Fang Zhao
National University of Singapore
elezhf@nus.edu.sg

Jianshu Li
National University of Singapore
jianshu@u.nus.edu

Tingfa Xu*
Beijing Institute of Technology
ciom_xtf1@bit.edu.cn

Jiashi Feng
National University of Singapore
elefjia@nus.edu.sg

ABSTRACT

Retrieving image content with a natural language expression is an emerging interdisciplinary problem at the intersection of multimedia, natural language processing and artificial intelligence. Existing methods tackle this challenging problem by learning features from the visual and linguistic domains *independently* while the critical semantic correlations bridging two domains have been underexplored in the feature learning process. In this paper, we propose to exploit sharable semantic attributes as "anchors" to ensure the learned features are well aligned across domains for better object retrieval. We define "attributes" as the common concepts that are informative for object retrieval and can be easily learned from both visual content and language expression. In particular, diverse and complex attributes (e.g., location, color, category, interaction between object and context) are modeled and incorporated to promote cross-domain alignment for feature learning from multiple perspectives. Based on the sharable attributes, we propose a deep Attribute-Preserving Metric learning (AP-Metric) framework that jointly generates unique query-sensitive region proposals and conducts novel cross-modal feature learning that explicitly pursues consistency over semantic attribute abstraction within both domains for deep metric learning. Benefiting from the cross-modal semantic correlations, our proposed framework can localize challenging visual objects to match complex query expressions within cluttered background accurately. The overall framework is end-to-end trainable. Extensive evaluations on popular datasets including ReferItGame [18], RefCOCO, and RefCOCO+ [43] well demonstrate its superiority. Notably, it achieves state-of-the-art performance on the challenging ReferItGame dataset.

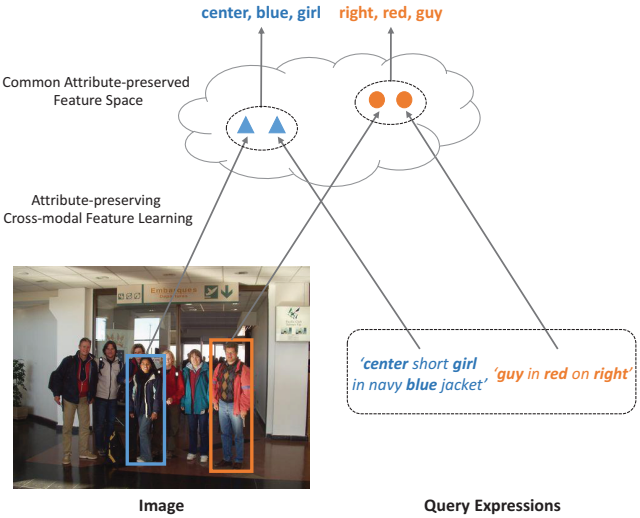


Figure 1: Demonstration of our motivation. Our proposed model learns to embed the coupled visual and linguistic features into a common attribute-preserving feature space where their consistency on abstracted semantic attributes is ensured. Thus the objects can be localized accurately by attribute-preserving metric learning across visual and linguistic domains.

KEYWORDS

Cross-modal; Object retrieval; Attribute

1 INTRODUCTION

Nowadays data are usually represented by various media modalities including image, text, audio, etc [22, 23]. Multimedia retrieval has been a heated research topic, such as image and text retrieval [3, 28, 37]. In this work, we consider a new multimedia retrieval protocol, i.e., natural language object retrieval. Given an image and a natural language expression, its goal is to localize the visual object described by the expression with an enclosed bounding box. Solving this problem not only requires visual and linguistic understanding of the image content and the query expression respectively, but also requires the search engine to understand the rich cross-modal

*Corresponding author.

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123439

correlations. Previous methods [16, 17] learn representation within the visual and the linguistic domain independently without considering the critical semantic correlations bridging the two domains, leading to unsatisfactory performance and poor user experience.

Although the query expression and the target image content lie in different domains, they share common semantic attributes of the object such as location, color, category and interactions with context, as they refer to the same semantics about the object with different expressions in the linguistic and the visual domain respectively. Such shared semantic attributes should be explored for promoting the consistency of cross-modal representations, which benefits the metric learning for cross-modal representation alignment. Specifically, traditional metric learning is not sufficient due to unaligned representations from different domains. The representations learned from two domains should not only contain sufficient and representative information of the target object, but also incorporate such information in a consistent way. We propose to preserve the shared semantic attributes bridging two domains during feature learning, such that the representations learned from both domains can keep representative and well aligned in the common feature space, which is critical for valid metric definition.

We propose a deep Attribute-Preserving Metric learning (AP-Metric) framework which embeds the visual and linguistic features into a common attribute-preserving feature space by enforcing attribute sharing constraints during feature learning for both image content and query expression, as shown in Figure 1. In particular, attribute sharing constraints are depicted by the attributes (e.g., "center", "blue", "girl") embedded in the natural language expression (e.g., "center short girl in navy blue jacket"). The proposed AP-Metric framework optimizes the feature embedding of image content and that of query expression by endowing both embedded representations the capability of accurately predicting such shared semantic attributes. Thus the learned cross-modal representation can preserve representative information of the target object and keep well aligned in the common feature space. In this way, the whole framework is a new cross-modal metric learning method that is able to align attributes from two different domains.

Specifically, the proposed AP-Metric framework first passes the image and query expression into a region-based visual encoding branch and a linguistic encoding branch to extract visual feature embeddings and linguistic feature embeddings, respectively. The visual encoding branch is constructed by a newly proposed query-sensitive region proposal network while the linguistic encoding branch is a recurrent Long-Short Term Memory (LSTM) [12] network. In order to embed the encoded features from two domains into a common feature space while preserving their shared semantic attributes of the target object (i.e., location, color, category, interactions with context), feature embeddings from two domains are assessed by the attribute-preserving cross-modal feature learning, which enables explicit learning of the attribute sharing constraints by predicting accurate attributes using the embedded cross-modal representation, through multi-label semantic attributes classification in the training phase. In addition, the contrastive loss is performed to narrow the distance between the embedded query features and the embedded visual features for the target image region and meanwhile enlarge that for the irrelevant image regions.

Moreover, we further exploit semantic correlations across visual and linguistic domains for region proposal generation. Previous methods [14, 43, 44] often ignore such correlations. They generally extract query-agnostic region proposals and then select the best matched one with the query expression as the retrieved result independently. Therefore they have two limitations: 1) proposal generation is independent of proposal ranking w.r.t. the linguistic information, leading to suboptimal solution; 2) proposals extracted by off-the-shelf region proposal algorithms (e.g., EdgeBoxes [47]) are usually irrelevant to the referred visual object, forming an efficiency bottleneck as a large number of proposals are required to achieve a satisfying proposal recall. An alternative is to train dedicated object detectors such as Fast R-CNN [8] and SSD [25] to extract proposals. However, it requires massive training data and predefined object categories of proposals, which is usually impractical in real-world natural object retrieval.

Intuitively, the query expression can provide useful cues for better covering and locating the referred object during region proposal generation. We develop a novel query-sensitive region proposal model to effectively incorporate the linguistic cues into the proposal generation. Specifically, taking the encoded query features and the image feature map as input, the proposed model learns to encode the correlations between the linguistic features and the visual representations at each spatial location on the image feature map to attend to the relevant regions w.r.t. the given query expression. Based on the produced encoded correlation maps, a set of proposals associated with their objectness scores are output for each spatial location on the image feature map, producing query-sensitive region proposals with higher recall and location accuracy compared to the traditional region proposal generation without considering linguistic cues.

We evaluate our method on the popular ReferItGame [18], RefCOCO, and RefCOCO+ [43] datasets, and experimental results show that it makes large improvements over previous methods with the same settings of training data, validating its superiority in both region proposal generation and final object retrieval.

To sum up, this work makes the following contributions. 1) We propose attribute-preserving metric learning for cross-modal retrieval by exploring the common semantic attribute abstraction bridging the visual and the linguistic domain. 2) We introduce a query-sensitive region proposal network which can generate query-sensitive region proposals w.r.t. specific linguistic information. 3) We are the first to embed the region proposal generation and cross-modal metric learning into a unified deep metric learning framework.

2 RELATED WORK

Cross-Modal Retrieval. Cross-modal retrieval has attracted much research attention due to explosive multimedia data. An important but difficult issue is to measure the content similarity between different data modalities including image and text [5, 31], text and audio [36], image and audio [21, 45]. One popular approach is to rely on manifold learning techniques [26, 42, 45, 46]. Zhang et al. [45] learned cross-modal correlations between visual and auditory feature spaces, and treated such correlations as complementary information for clustering on image-audio datasets. Mahadevan et

al. [26] proposed maximum covariance unfolding (MCU), a manifold learning algorithm for simultaneous dimensionality reduction of data from different input modalities. Another approach is to learn correlations between modalities [21, 40]. Canonical Correlation Analysis (CCA) is one of the most popular subspace learning methods for establishing inter-modal relationships between different modalities of data, which has been widely used for cross-media retrieval [9, 29, 32]. Andrew et al. [1] presented a nonlinear extension of CCA, Deep Canonical Correlation Analysis (DCCA), which is a deep learning method to learn complex nonlinear transformations of data from different modalities such that the resulting representations are highly linearly correlated.

Grounding Objects from Image Descriptions. The methods of grounding objects from image descriptions take an image and its description sentence as input, and align sentence fragments to image regions. Karpathy et al. [17] proposed a model to learn a multi-modal embedding space for fragments of images and sentences and to reason about their latent, inter-modal alignment. [16] proposed a bidirectional recurrent neural network to compute word representations in the sentences and aligned sentence snippets to the visual regions described through a multi-modal embedding. Plummer et al. [30] used CCA to learn a shared semantic space to associate phrases in image descriptions and image regions. [20] used a structured prediction model to estimate the text-to-image alignment and reasoned about object co-reference in text for 3D scene parsing. Rohrbach et al. [34] used a recurrent network to encode the phrase and then learned to attend to the relevant image region by trying to reconstruct the input phrase.

Natural Language Object Retrieval. Natural language object retrieval localizes a target object within a given image based on a natural language query of the object. The work [11] first mapped the candidate object regions to sets of words using several learned image-to-text projections, and then compared the natural language query with the sets of words predicted for each candidate region and selected the best match. Hu et al. [14] used a recurrent neural network model to learn a scoring function based on the text query, candidate regions, their spatial configurations and global context, and took the candidate region with the highest score as the retrieved result. [27] scored the phrase on a set of proposals using a caption generation framework to select the proposal with the highest probability. Yu et al. [43] proposed to incorporate visual comparison based context into referring expression models, showing advantages in both referring expression generation and comprehension. Wu et al. [41] initialized a bounding box to cover the whole image, and trained an agent with deep reinforcement learning, which learns to move and reshape the bounding box to localize the object according to the referring expression.

3 OUR MODEL

Figure 2 shows the overall architecture of the proposed AP-Metric framework, which mainly includes an LSTM-based linguistic feature encoder, a region-based visual feature encoder constructed by a query-sensitive region proposal network, two modal-specific feature embedding sub-networks and an attribute-preserving cross-modal feature learning module. Specifically, in the training phase, given an image and a query expression, a recurrent LSTM-based linguistic feature encoder is first applied to encode the query expression to a fixed-length feature vector. The input image is fed into several convolutional layers and pooling layers to extract its spatial feature map. Then taking the encoded query features and the image feature map as input, a query-sensitive region proposal network is proposed to produce region proposals w.r.t. the given query expression. Next, region descriptors are extracted for the generated proposals using Region-of-Interest (RoI) pooling [8] on top of the image feature map, which are forwarded into several fully-connected layers to produce encoded visual region features combined with the location information for each proposal. To ensure the visual and linguistic features are well aligned, the visual and linguistic feature embedding sub-networks embed the encoded visual and linguistic features into our proposed common attribute-preserving feature space separately by ensuring their consistency on semantic attribute abstraction, guided by the novel attribute-preserving cross-modal feature learning. Moreover, the contrastive loss is applied to conduct metric learning for the alignment of the embedded visual and linguistic features in the common feature space. During testing, we measure the distance between the embedded features of the given query expression and those of each generated region proposal, and take the one whose embedded region features have the smallest distance to the embedded query features as the retrieved result.

3.1 Linguistic and Visual Feature Encoding

3.1.1 LSTM-based Linguistic Feature Encoder. Given a query expression describing an object in the image, the linguistic feature encoder aims to encode it into a fixed-length feature vector to facilitate subsequent retrieval and localization. Following the convention in natural language processing, we first represent each word in the query expression as a one-hot vector and then embed it into the semantic space through a linear word embedding transformation as in [14]. Then the sequence of embedded word feature vectors is input to a recurrent Long-Short Term Memory (LSTM) [12] network with D_l dimensional hidden states. Here we set D_l as 1000 in our implementation. We use LSTM to encode the sequence of word vectors as it has been proved effective in many language modeling tasks [7, 38, 39]. Denote the query expression with T words as $S = (w_1, ..., w_T)$, where w_t represents the embedded vector from the embedding matrix for the word t. At each time step t, the LSTM takes w_t as input. In this way, we encode the query expression of an arbitrary length word by word with an LSTM, and obtain an encoded vector representation of the query using the hidden state h_T at the final time step T as

$$h_T = f_{LSTM}(S). \tag{1}$$

3.1.2 Region-based Visual Feature Encoder. Given an image and a set of region proposals, the region-based visual feature encoder aims to generate a representative feature representation of the visual content for each region proposal. We first feed the whole image into several convolutional layers and pooling layers to extract its spatial feature map. In particular, we use the VGG16 architecture [35] as in [14, 34] in our implementation. The resulting spatial feature map, denoted as "Conv5_3", has the dimension of $D_v \times h \times w$. Here $h = H/s$ and $w = W/s$, where H and W denote the height and

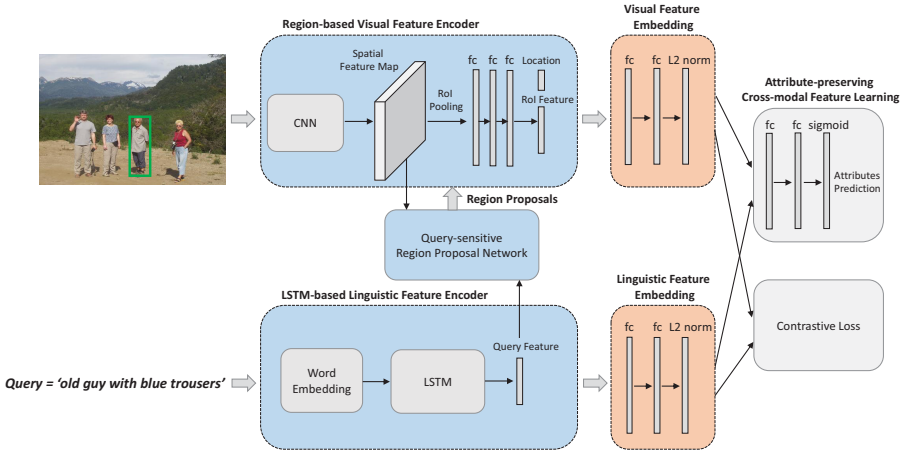


Figure 2: Overall architecture of the proposed AP-Metric framework. Taking an image and a query expression as input, the model first encodes the query expression through an LSTM-based linguistic feature encoder. The region-based visual feature encoder generates encoded visual features based on the region proposals from the query-sensitive region proposal network, which produces query-sensitive region proposals by incorporating both visual and linguistic information. Then, two modal-specific feature embedding sub-networks are utilized to embed the encoded visual and linguistic features into a common attribute-preserving feature space through attribute-preserving cross-modal feature learning, facilitating cross-modal feature alignment in the contrastive loss optimization.

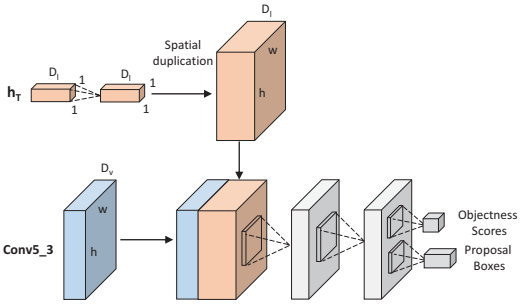


Figure 3: Architecture of the query-sensitive region proposal network. Taking the encoded query features h_T and image feature map "Conv5_3" as input, the network first transforms h_T using a 1×1 convolution and then duplicates it spatially to concatenate with "Conv5_3" along the channel axis, forming a combined feature map, on top of which a 1×1 convolution is utilized to encode the correlations between the linguistic and visual features at each spatial location. Finally, a 3×3 convolution followed by two sibling 1×1 convolution is implemented to output objectness scores and regressed proposal boxes for each spatial location.

width of the input image respectively, and $s = 16$ represents the pixel stride on the last convolutional layer of the VGG16 model.

Then, we use RoI pooling [8] on top of the last spatial feature map to extract a feature descriptor of dimension $D_v \times 7 \times 7$ for each region proposal. Here D_v equals 512 as defined in the VGG16 model. The generated feature descriptor is then fed into three fully-connected layers, i.e., "fc6" and "fc7" of the VGG16 model and a newly added layer, to transform it into a fixed-length feature vector. As the units on the last spatial feature map produced by VGG16 model have a very large receptive field [2], the produced feature vector for each region proposal has the potential to aggregate contextual information from nearby regions, which is beneficial for reasoning about interaction between visual objects.

For better representing the image regions, we further incorporate the spatial location information for each region proposal into feature encoding. For a specific proposal, we represent its spatial configuration as a feature vector:

$$f_{spatial} = [w, h, x_1, y_1, x_2, y_2, x_c, y_c], \tag{2}$$

where (w, h) specifies the width and the height of the region proposal, and (x_1, y_1), (x_2, y_2) and (x_c, y_c) denote the coordinates of its top-left corner, bottom-right corner and center point, respectively. We normalize the width and the height of the image as 2 and set the coordinates of the image center as $(0, 0)$. We concatenate the region feature vector and the spatial configuration for each region proposal together, forming a compact feature representation containing both visual and location information.

3.2 Query-sensitive Region Proposal Network

Based on the produced image feature map and encoded query features, the query-sensitive region proposal network aims to generate a set of region proposals matching the given query expression, each of which is also associated with a predicted objectness score. This is substantially different from traditional query-blind proposal generation methods. Exploiting the semantic correlations between linguistic meanings and visual representation at each spatial location can explicitly improve the recall and location accuracy of the region proposals. Moreover, the proposal generation and ranking process w.r.t. the given query expression can be optimized jointly and benefit each other through a shared feature representation.

In order to output a set of region proposals associated with their predicted objectness scores at each spatial location on the image feature map based on its correlation with the given linguistic information, fully convolutional classifiers over the image feature

map and the encoded query features are utilized. Specifically, as shown in Figure 3, the network takes as input the encoded query feature h_T for the given query expression and the image's last spatial feature map "Conv5_3" of dimension $D_v \times h \times w$. To attend to the regions on the spatial feature map which are most relevant to the query expression, we first replicate h_T to the same size of the image feature map spatially to obtain a $D_l \times h \times w$ feature blob, which is then concatenated with the image feature map along the channel axis, forming a combined feature map of dimension $(D_v + D_l) \times h \times w$ containing both visual and linguistic information. On top of the combined feature map, a 1×1 convolution is implemented to encode the correlations between the query features and the visual features at each spatial location of the image feature map, to attend to the relevant regions to the given query expression.

Next, following the Region Proposal Network (RPN) proposed in [33], a 3×3 convolution is implemented to slide over the produced correlation feature maps, followed by two sibling 1×1 convolutional layers — a box-classification layer and a box-regression layer to output predicted objectness scores and regressed proposal boxes for each sliding location, respectively. To generate reference boxes, we use 4 scales and 3 aspect ratios, yielding $k = 12$ anchors as defined in [33] at each sliding position. We minimize an objective function following the multi-task loss (denoted as L_{rpn}) as adopted in [33], which includes a classification loss and a bounding box regression loss, to optimize the parameters of the proposed query-sensitive region proposal network.

3.3 Attribute-preserving Metric Learning

Given encoded features for the image regions and query expression, the proposed model further learns to embed them into a common feature space for attribute-preserving metric learning. Specifically, we embed the encoded visual and linguistic features into a common attribute-preserving feature space found by two modal-specific embedding sub-networks, each of which consists of two fully-connected layers followed by an L2 normalization layer. By enforcing attribute sharing constraints during the feature embedding through a novel attribute-preserving cross-modal feature learning approach, the common attributes shared by both encoded visual and linguistic features are preserved in the transformed common feature space, facilitating the cross-modal feature alignment in the following contrastive loss optimization.

3.3.1 Attribute-preserving Cross-modal Feature Learning. Although the encoded features for the target image region and query expression are from the visual and the linguistic domain respectively, they undoubtedly share common semantic attributes of the same referred object. We propose to preserve such shared semantic attribute abstraction in the common feature space by enforcing attribute sharing constraints during feature embedding, which is formulated as encouraging the embedded features from both domains to predict attributes accurately simultaneously, through multi-label semantic attributes classification.

Specifically, we find that location, color, category and interactions with other objects are among the most representative attributes for object localization. Thus, we define in total R attribute categories which are divided into five groups. The first three groups are spatially relevant and each group contains 2 binary attributes indicating whether the target object resides at the left/right, top/bottom, and center (middle)/corner of the image respectively. The fourth group contains L binary attributes indicating the color of the target object. The fifth group contains C binary attributes indicating the categories of the target object and those interacting with the target object as described in the query expression. Considering efficiency and training samples limitation, we do not take all possible colors and object categories into consideration. In our implementation, we set L and C as 10 and 74 respectively, forming $R = 90$ binary attributes totally. We select the top 10 colors and the top 74 object categories according to their statistical frequencies in the query expressions on the whole training set.

For generating labels for the location and color binary attributes, we set each location and color label (1 indicates true while 0 indicates false) according to the description of the target object in the given query expression. For category attributes label generation, given a query expression, if it contains any vocabulary (w.r.t either the target or the environmental objects) among the predefined C categories, we set the label corresponding to that category to be 1, otherwise 0. If none of the predefined attributes in a specific group can be inferred from the given query expression, we set all the attribute labels in that group to be -1 so that no loss will be produced in the attribute group for this instance during training (see below for loss definition). Thus we formulate the semantic attributes of the target object as well as the categories of its interacted objects for contextual relationship representation in a simple yet effective way.

We preserve the common semantic attributes during feature embedding by endowing the embedded features from both domains with the capability of predicting correct semantic attributes. Specifically, we extract the embedded features of the query expression and the target image region from the second fully-connected layer of the linguistic and the visual feature embedding sub-network, respectively, and feed both features into two fully-connected layers followed by a sigmoid layer to output discrete probability estimates for each of the R binary attributes. Thus we formulate the semantic attributes prediction as a multi-label classification task by minimizing the sigmoid cross entropy loss as

$$L_{attr} = -\sum_{l=1}^{R}[p_l \log \hat{p}_l + (1 - p_l)\log(1 - \hat{p}_l)], \tag{3}$$

where $\hat{p}_l$ and p_l are the prediction and the target for label l, respectively. In this way, both the embedded visual and linguistic features are enforced to preserve the representative information, i.e., the semantic attributes of the target object accurately during feature learning. By keeping the consistency in such semantic attributes, the embedded features from two domains can be well aligned in the common attribute-preserving feature space.

3.3.2 Attribute-preserving based Metric Learning. We conduct metric learning using the embedded visual and linguistic features in the common attribute-preserving feature space, which aims to narrow the distance to the embedded query features from embedded visual features of the target image region and meanwhile enlarge such distance from embedded visual features of the irrelevant image regions. In our implementation, we randomly select $N = 64$ training samples from the region proposals generated by the query-sensitive region proposal network for each image, in order to generate embedded region features to conduct metric learning with

Table 1: Accuracy on ReferItGame dataset.

Method	Accuracy
LRCN [6]	8.59%
CAFFE-7K [11]	10.38%
SCRC [14]	17.93%
GroundeR [34]	28.51%
Wu et al. [41]	36.18%
Ours	**44.18%**

Table 2: Accuracy on RefCOCO dataset. All methods use only RefCOCO training set for training.

RefCOCO			
Method	Test A	Test B	Validation
SCRC [14]	18.47%	20.16%	19.02%
Wu et al. [41]	54.78%	41.58%	48.19%
Ours	**55.21%**	**46.22%**	**52.35%**

the embedded query features. We take 50% of the samples from region proposals that have intersection over union (IoU) of at least 0.5 with the ground truth bounding box. These samples are taken as positive training samples and labeled with $y = 1$. The remaining samples are selected from region proposals that have IoU of less than 0.5 with the ground truth bounding box, which are taken as negative training samples and labeled with $y = 0$.

We aim to encourage the distance between the embedded query features and the embedded visual features for the positive samples to be zero, while that for the negative samples to be larger than some enforced margin m. To this end, we use the contrastive loss defined:

$$L_{cont} = \frac{1}{2N}\sum_{n=1}^{N}(y_n)d_n^2 + (1-y_n)\max(m-d_n,0)^2, \quad (4)$$

where

$$d_n = \|q - v_n\|_2, \quad (5)$$

where q denotes the embedded query features, and v_n and y_n represent the embedded visual features and the label for the $n-th$ selected training sample, respectively. We set m as 3 in our implementation. Traditional metric learning only measures the distance between the embedded features without considering the critical semantic correlations bridging two domains. Through attribute-preserving metric learning, embedded cross-modal representation can be well aligned in advance, which is critical for valid metric definition.

Denote θ as the parameters of the whole network, we obtain θ by optimizing the overall loss function which is formulated as a multi-task learning problem:

$$\theta = \arg\min(L_{cont} + \alpha L_{attr} + \beta L_{rpn}), \quad (6)$$

where α and β represent the balancing parameters for different loss functions, set to be 0.2 and 1.0 in our implementation.

During testing, given an input image and a query expression, we compute the distance between the embedded query features and the embedded visual features for the region proposals generated by the query-sensitive region proposal network, and the region proposals with the smallest distance are taken as the retrieved results.

Table 3: Accuracy on RefCOCO+ dataset. All methods use only RefCOCO+ training set for training.

RefCOCO+			
Method	Test A	Test B	Validation
SCRC [14]	14.43%	13.25%	13.72%
Wu et al. [41]	40.39%	22.81%	31.93%
Ours	**45.20%**	**32.24%**	**40.19%**

Table 4: Comparisons of accuracy with several model variants on ReferItGame dataset.

Method	Accuracy
Ours (RPN)	37.94%
Ours (w/o attributes)	40.10%
Ours (w/o spatial)	40.65%
Ours	**44.18%**

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

4.1.1 ReferItGame Dataset. The ReferItGame dataset [18] contains 20,000 images from IAPR TC-12 dataset [10], and over 99,000 image regions associated with natural language expressions. We only use the bounding boxes of annotated regions provided by [14] during training and evaluation. Following [14], the whole dataset is split into two halves at the image level for training and testing respectively. We construct image-bounding box-description tuples on all annotated image regions as training instances, leading to 59,976 instances in the trainval set and 60,105 in the test set. We train our model on the trainval set. For evaluation, we compute the accuracy as the ratio of queries for which the retrieved box overlaps with the ground truth bounding box by at least 0.5 IOU.

4.1.2 RefCOCO and RefCOCO+ Dataset. The RefCOCO dataset and the RefCOCO+ dataset [43] are collected on MSCOCO images [24], using the ReferItGame [18]. The RefCOCO dataset consists of 142,209 referring expressions for 50,000 objects in 19,994 images, and the RefCOCO+ dataset has 141,564 referring expressions without location words for 49,856 objects in 19,992 images. We use the original split provided by each dataset. Both datasets provide person vs. object splits for evaluation. The images in TestA contain multiple people while images in TestB contain multiple objects of other categories. For evaluation, we also compute the accuracy as above mentioned.

4.2 Implementation Details

The implementation is based on the Caffe platform [15]. We use the VGG16 model [35] pre-trained on ImageNet [4] for initialization. The parameters of all other newly added layers including the LSTM unit, word embedding layer, convolutional layers, and fully connected layers are initialized from zero-mean Gaussian distributions with standard deviation 0.01. We use images in each dataset as input without resizing operation. The whole network is optimized using Adam [19] with a fixed learning rate 0.0001 on a single NVIDIA GeForce GTX TITAN X GPU with 12GB memory. During testing, for fair comparison with SCRC [14] which uses 100 top scoring proposals from EdgeBoxes [47], we also select 100 top scoring region proposals generated by the query-sensitive region proposal network per image on test and validation sets.

Figure 4: Examples from the test set of ReferItGame dataset. The ground-truth bounding boxes of objects are annotated with red rectangles. The green rectangles and yellow rectangles represent the retrieved results by the proposed AP-Metric framework and the model variant without performing attribute-preserving metric learning. Best viewed in color.

4.3 Performance Comparison

4.3.1 ReferItGame Dataset. Table 1 provides the comparisons of our method with other state-of-the-arts in accuracy on ReferItGame dataset [18]. It can be observed that our method achieves the highest accuracy of 44.18%, outperforming all the baseline methods [6, 11, 14, 34, 41] by a large margin. Particularly, the proposed method improves the performance significantly over SCRC [14] by 26.25%. Similarly, it outperforms the previous state-of-the-art method of Wu et al. [41] by 8.00%, which well demonstrates its superiority.

Figure 4 shows some correctly retrieved object examples (green rectangles) from the ReferItGame test set, where the highest scoring region proposal matches the ground truth. One can see that our method can correctly localize the referred objects given an image with different natural language queries. We also show the retrieved results (yellow rectangles) by the model variant without conducting attribute-preserving cross-modal feature learning. It can be observed that the proposed AP-Metric framework can provide more accurate localization by taking advantage of preserved semantic attributes (e.g., category, location, and color) embedded in the natural language expression (e.g., "light left", "bush to the left", and "middle blue kid"), while the model variant without attribute-preserving metric learning fails.

4.3.2 RefCOCO and RefCOCO+ Dataset. For the RefCOCO dataset and the RefCOCO+ dataset [43], we use only the original training set of each dataset for training, and test our method on the test and the validation set of the two datasets respectively. We compare the results of our method with those of SCRC [14] and Wu et al. [41], which also use no extra labeled data for training. For fair comparison, we do not include the results of [43] and [44] as they pre-train dedicated object detectors using massive extra training data, i.e., the validation set and the trainval set of MSCOCO [24].

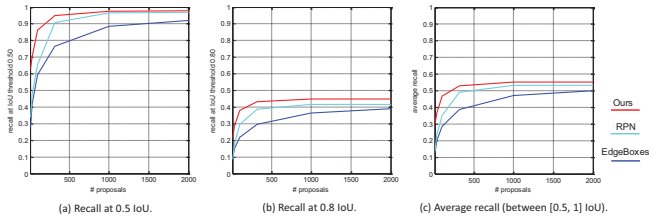
As shown in Tables 2 and 3, the proposed method outperforms the methods of SCRC and [41] on both datasets by a large margin. Specifically, on the RefCOCO dataset, our method makes a large improvement over SCRC by 36.74%, 26.06%, and 33.33% on the TestA, TestB and validation set respectively. It also outperforms the method of [41]: 55.21% vs. 54.78%, 46.22% vs. 41.58% and 52.35% vs. 48.19% on the three subsets respectively. Similarly, on the RefCOCO+ dataset, the proposed method makes a large improvement of 30.77%, 18.99%, and 26.47% compared to SCRC, and 4.81%, 9.43%, and 8.26% compared to [41] on the TestA, TestB and validation set respectively.

4.4 Ablation Studies

We investigate the effectiveness of different components of our model through experiments on the ReferItGame dataset.

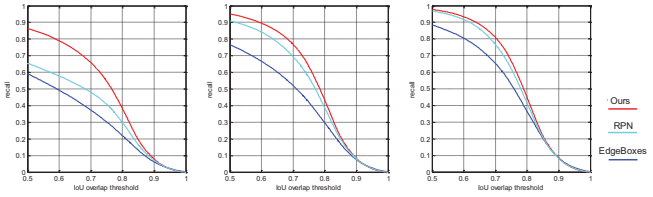
4.4.1 Effectiveness of Query-sensitive Region Proposal Network. The proposed query-sensitive region proposal network learns to generate query-sensitive region proposals by incorporating the encoded linguistic features of each query expression. In order to verify its advantage, we compare our model with the variant where the query-sensitive region proposal network is replaced with Region Proposal Network (RPN) [33] to generate general region proposals without considering the specific linguistic information, denoted as "Ours (RPN)".

We first investigate the region proposals generated by both networks on the ReferItGame test set. In addition, we also compare with those extracted by EdgeBoxes [47]. As shown in Figures 5 and 6, the proposed query-sensitive region proposal network outperforms EdgeBoxes by a large margin in terms of proposal recall, demonstrating its superiority over the off-the-shelf region proposal algorithms. Figure 5 (a) and (b) show the proposal recall vs. the number of proposals for different fixed IoU thresholds. It can be observed that the proposed query-sensitive region proposal network provides region proposals with much higher recalls compared to those generated by RPN [33], especially when given a more strict IoU threshold (e.g., 0.8), validating that the incorporated linguistic information can provide beneficial cues for accurately localizing the target objects. Figure 5 (c) presents the average recall between [0.5, 1] IoU vs. the number of proposals, which summarizes proposal performance across IoU thresholds [13]. One can observe that



(a) Recall at 0.5 IoU. (b) Recall at 0.8 IoU. (c) Average recall (between [0.5, 1] IoU).

Figure 5: Recall vs. IoU threshold comparisons of our query-sensitive region proposal network, RPN [33] and EdgeBoxes [47] on ReferItGame test set.



(a) 100 proposals per image. (b) 500 proposals per image. (c) 2000 proposals per image.

Figure 6: Recall vs. number of proposals comparisons of our query-sensitive region proposal network, RPN [33] and EdgeBoxes [47] on ReferItGame test set.

the proposed method keeps a higher recall compared to RPN across the entire range of region proposal number. We further evaluate the proposal recall across the IoU threshold range of [0.5, 1] for different fixed numbers of region proposals, as shown in Figure 6 (a), (b) and (c). Our method shows higher recall consistently compared to RPN for different numbers of region proposals, validating the superiority of the query-sensitive region proposals.

Figure 7 shows some examples of the generated region proposals by the proposed query-sensitive region proposal network, RPN [33], and EdgeBoxes [47]. For clear observation, we select 20 top scoring region proposals generated from each method. It can be observed that the region proposals generated by our method can localize and cover the objects described by the given query expressions more accurately compared to those produced by RPN and EdgeBoxes, validating the effectiveness of incorporating linguistic cues for query-sensitive region proposal generation.

A similar conclusion can be drawn based on the improvement in the retrieval accuracy. As can be seen in Table 4, our model improves the accuracy by 6.24% compared to "Ours (RPN)", which verifies that the proposed query-sensitive region proposal network can generate better region proposals w.r.t. the query expression with high recall and location accuracy, and improve the subsequent proposal ranking process benefited from a shared feature representation.

4.4.2 Effectiveness of Attribute-preserving Cross-modal Feature Learning. The proposed method performs attribute-preserving metric learning through predicting accurate attributes using the embedded features from the visual and the linguistic domains. In order to verify the effectiveness of preserving common semantic attributes abstraction bridging two domains in the common feature space, we compare the results of our model with the variant where no attributes sharing constraints are performed during feature learning, denoted as "Ours (w/o attributes)". Table 4 shows that our model increases the accuracy by 4.08% compared to "Ours (w/o attributes)", showing that the model can better exploit the connections between

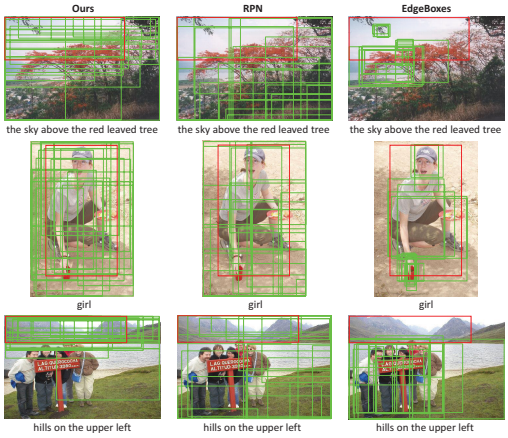


Figure 7: Comparisons of generated region proposals on ReferItGame test set. The three columns show the generated region proposals by the proposed query-sensitive region proposal network, RPN [33], and EdgeBoxes [47], respectively. The ground-truth bounding boxes of the referred objects and the generated top scoring region proposals are represented as red and green rectangles, respectively.

the visual and linguistic features by ensuring their consistency on semantic attribute abstraction.

4.4.3 Effectiveness of Spatial Location Information. During visual feature encoding, we combine the spatial location information for each region proposal with the extracted region features together, forming an enhanced feature representation. To analyze the advantage of incorporating spatial location information, the results of the variant that generates encoded visual feature representation without considering spatial location information are reported, i.e., "Ours (w/o spatial)" in Table 4. It can be observed that "Ours (w/o spatial)" decreases the accuracy by 3.53% compared to our model, verifying that incorporating the spatial location information for each region proposal into visual feature encoding can provide beneficial cues for better locating the target objects.

5 CONCLUSION

In this paper we propose an end-to-end trainable deep framework for jointly generating query-sensitive region proposals and performing attribute-preserving metric learning for natural language object retrieval. Particularly, we encode and embed the image content and query expression into a common attribute-preserving feature space by keeping their consistency on semantic attribute abstraction to enable attribute-preserving based metric learning across visual and linguistic domains. Moreover, a query-sensitive region proposal network is designed to extract specific region proposals w.r.t the query expression with high recall and location sensitivity. Extensive experiments have well demonstrated the superiority of the proposed method.

ACKNOWLEDGEMENT

This work was partially supported by China Scholarship Council (Grant No. 201506030045). The work of Jiashi Feng was partially supported by NUS startup R-263-000-C08-133, MOE Tier-I R-263-000-C21-112 and IDS R-263-000-C67-646.

REFERENCES

[1] Galen Andrew, Raman Arora, Jeff A Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 1247–1255.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* (2014).

[3] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)* 40, 2 (2008), 5.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[5] Ludovic Denoyer and Patrick Gallinari. 2004. Bayesian network model for semi-structured document classification. *Information Processing & Management* 40, 5 (2004), 807–827.

[6] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2625–2634.

[7] Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. 2016. Contextual LSTM (CLSTM) models for Large scale NLP tasks. *arXiv preprint arXiv:1602.06291* (2016).

[8] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 1440–1448.

[9] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision* 106, 2 (2014), 210–233.

[10] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. [n. d.]. The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems. In *OntoImage 2006 Workshop on Language Resources for Content-based Image Retrieval during LREC 2006 Final Programme*.

[11] Sergio Guadarrama, Erik Rodner, Kate Saenko, Ning Zhang, Ryan Farrell, Jeff Donahue, and Trevor Darrell. 2014. Open-vocabulary Object Retrieval. In *Robotics: Science and Systems*, Vol. 2. Citeseer, 6.

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[13] Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. 2016. What makes for effective detection proposals? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 4 (2016), 814–830.

[14] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4555–4564.

[15] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM, 675–678.

[16] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.

[17] Andrej Karpathy, Armand Joulin, and Fei Fei F Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems*. 1889–1897.

[18] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *The Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 787–798.

[19] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[20] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3558–3565.

[21] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K Sethi. 2003. Multimedia content processing through cross-modal association. In *Proceedings of the 11th ACM International Conference on Multimedia*. ACM, 604–611.

[22] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. 2017. Recurrent Topic-Transition GAN for Visual Paragraph Generation. *arXiv preprint arXiv:1703.07022* (2017).

[23] Xiaodan Liang, Lisa Lee, and Eric P Xing. 2017. Deep variation-structured reinforcement learning for visual relationship and attribute detection. *arXiv preprint arXiv:1703.03054* (2017).

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 740–755.

[25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. SSD: Single shot multibox detector. In *European Conference on Computer Vision*. Springer, 21–37.

[26] Vijay Mahadevan, Chi W Wong, Jose C Pereira, Tom Liu, Nuno Vasconcelos, and Lawrence K Saul. 2011. Maximum covariance unfolding: Manifold learning for bimodal data. In *Advances in Neural Information Processing Systems*. 918–926.

[27] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11–20.

[28] Charles T Meadow, Bert R Boyce, Donald H Kraft, and Carol Barry. 2007. *Text Information Retrieval Systems*. Academic Press.

[29] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2014. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3 (2014), 521–535.

[30] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*. 2641–2649.

[31] Guo-Jun Qi, Charu Aggarwal, and Thomas Huang. 2011. Towards semantic knowledge propagation from text corpus to web images. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, 297–306.

[32] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, 251–260.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. 91–99.

[34] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*. Springer, 817–834.

[35] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[36] Malcolm Slaney. 2002. Semantic-audio retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 4. IV–4108.

[37] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 12 (2000), 1349–1380.

[38] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM Neural Networks for Language Modeling. In *Interspeech*. 194–197.

[39] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*. 3104–3112.

[40] Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. 2002. Inferring a semantic representation of text via cross-language correlation analysis. In *Advances in Neural Information Processing Systems*, Vol. 1. 4.

[41] Fan Wu, Zhongwen Xu, and Yi Yang. 2017. An End-to-End Approach to Natural Language Object Retrieval via Context-Aware Deep Reinforcement Learning. *arXiv preprint arXiv:1703.07579* (2017).

[42] Yi Yang, Dong Xu, Feiping Nie, Jiebo Luo, and Yueting Zhuang. 2009. Ranking with local regression and global alignment for cross media retrieval. In *Proceedings of the 17th ACM International Conference on Multimedia*. ACM, 175–184.

[43] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*. Springer, 69–85.

[44] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2016. A Joint Speaker-Listener-Reinforcer Model for Referring Expressions. *arXiv preprint arXiv:1612.09542* (2016).

[45] Hong Zhang, Yueting Zhuang, and Fei Wu. 2007. Cross-modal correlation learning for clustering on image-audio dataset. In *Proceedings of the 15th ACM International Conference on Multimedia*. ACM, 273–276.

[46] Yue-Ting Zhuang, Yi Yang, and Fei Wu. 2008. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia* 10, 2 (2008), 221–229.

[47] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*. Springer, 391–405.