Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-15779-y

Multimodal query-guided object localization

Aditay Tripathi[1] · **Rajath R Dani**[1] · **Anand Mishra**[2] · **Anirban Chakraborty**[1]

Received: 3 December 2022 / Revised: 26 March 2023 / Accepted: 2 May 2023

Abstract
Recent studies have demonstrated the effectiveness of using hand-drawn sketches of objects as queries for one-shot object localization. However, hand-drawn crude sketches alone can be ambiguous for object localization, which could result in misidentification, e.g., a sketch of a laptop could be confused for a sofa. To overcome this, we propose a novel multimodal approach to object localization that combines sketch queries with linguistic category definitions, allowing for a better representation of visual and semantic cues. Our approach employs a cross-modal attention scheme that guides the region proposal network to obtain relevant proposals. Further, we propose an orthogonal projection-based proposal scoring technique that effectively ranks proposals with respect to the query. We evaluated our method using hand-drawn sketches from the 'Quick, Draw!' dataset and glosses from 'WordNet' as queries on the widely-used MS-COCO dataset, and achieve superior performance compared to related baselines in both open- and closed-set settings.

Keywords Sketch · Open-set object localization · Gloss · Cross-modal localization · Cross-modal attention

1 Introduction

We have seen breakthroughs in object detection literature in the last decade, and it is partly due to the advancements in deep learning [36, 40, 59, 60]. However, most of these successful models are still limited to 'closed-world' settings, where the object localization and classi-

✉ Anirban Chakraborty
anirban@iisc.ac.in

Aditay Tripathi
aditayt@iisc.ac.in

Rajath R Dani
rajathrdani@gmail.com

Anand Mishra
mishra@iitj.ac.in

[1] CDS, Indian Institute of Science, 560012 Bengaluru, Karnataka, India

[2] CSE, Indian Institute of Technology, 342037 Jodhpur, Rajasthan, India

Published online: 11 July 2023

fication tasks are limited to a predefined set of categories whose examples are used during the training phase. In this work, we study a more challenging task of open-set query-guided object localization with the following goal – given an image of a natural scene and an object query, localize all the instances of the queried object in the image, even if no sample for this queried object is assumed available during the training phase. In the literature, query-guided object localization has been attempted using either object category name [79] or an image of the object as a query [25, 67]. However, it is possible to encounter scenarios where neither an image of the object nor the class label is available as a query. Such a scenario can arise (i) due to privacy reasons or (ii) when the object of interest is uncommon (not a natural object, e.g., parts of a machine). However, even in such a scenario, it is often easy to find a crude drawing or natural language description of the query object. We, therefore, want to explore the following task-*Can a hand-drawn sketch or natural language description of any object be used for localizing all the instances of the corresponding object in a natural scene?* We investigate the answer to this research question in this work.

In our earlier work [71], we introduced the novel idea of using hand-drawn sketches of objects as queries towards localizing objects in a natural scene. Sketches provide an abstract visual representation of the objects. Most free-hand sketches (e.g., sketches in Quick Draw) lack serious visual content, such as the appearance, color, and texture of the drawn objects. Often, these sketches only provide noisy outlines depicting the global shapes of an object and lack any finer structural details. For example, consider Fig. 1, where a hand-drawn sketch of a laptop might be ambiguous for object localization as it could be confused for a sofa. These, understandably, lead to very limited success in the sketch-guided object localization task. On the contrary, by combining different modalities, such as visual (in the form of a hand-drawn sketch of the query object) and text (in the form of a natural language description of the query object), it is possible to leverage complementary and intricate details on an object's shape, appearance or texture and sometimes even the semantic relationship of the query with other objects in the scene. Judiciously combining these modalities may yield a much richer representation of the query with less ambiguity and potentially lead to a better open-set localization performance. In this work, in addition to a sketch query, we use a linguistic

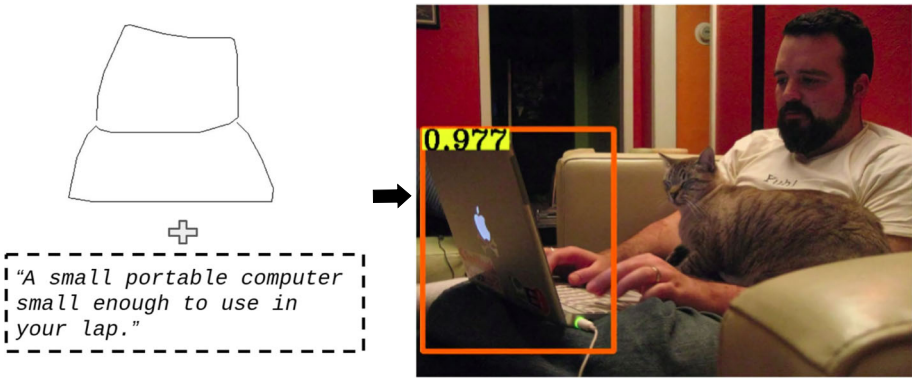

Fig. 1 Given an image and a query, our aim is to localize the object in the image (a laptop in this example). A hand-drawn sketch of a laptop alone, when used as a query, might be ambiguous for object localization as it could be confused for a sofa. On the other hand, descriptions obtained from different modalities such as a category label, e.g. "laptop" or a linguistic definition of the category, e.g., "a small portable computer small enough to use in your lap" along with the sketch query, give better visual and semantic cues for the object localization

definition of the category also known as gloss, e.g. “a small portable computer small enough to use in your lap" as multimodal queries for the object localization task.

There are several technical challenges associated with multimodal query-guided object localization, such as (i) a large domain gap between the query modalities (e.g., text, sketches, etc.) and the target natural images, and (ii) diverse and minimal information present in queries. For example, a sketch query captures abstract shape information of an object, whereas a text query often captures partial semantic information about the object category. In order to address these challenges, one plausible solution is to use the standard region proposal network (RPN) and score the generated proposals against the query. However, the standard RPN does not utilize query information to generate region proposals; therefore, the relevant region proposals may not even be generated, especially for the open set case. On the contrary, in our framework, we propose a cross-modal attention scheme towards generating object proposals relevant to the input queries. A preliminary version of the same was proposed in our earlier conference paper [71]. The novel extended version of the cross-modal attention strategy is designed to generate a spatial compatibility matrix by comparing the combined query, i.e., concatenated sketch and text representation, with the local image representations obtained from each image feature map location, thereby incorporating query information during proposal generation. In other words, our proposal generation step is query-aware. A unique advantage of this strategy is that it enables the generation of proposals, even for those object categories that are unseen during training. Further, we propose a novel multimodal proposal scoring scheme to score object proposals with features from multiple modalities. The proposed scheme first defines a subspace constructed using queries as the basis vectors, then the feature vector of each proposal is projected onto this subspace. Finally, the projected vector is utilized to score each of the proposals. By being an orthogonal projection, the proposed scheme generates a vector in the subspace of the queries which is closest to the object proposal vector, and hence it leads to better scoring between the queries and proposals. Moreover, the proposed scoring scheme is able to capture complementary information present in multiple modalities that enables it to achieve superior performance for open-set object localization.

We have performed extensive experiments on multimodal query-guided object localization on public benchmarks. We show results for both the open-set, i.e., disjoint train-test categories, and the closed-set, i.e., common train-test categories settings, and perform extensive ablation studies. Our method with sketch and gloss as composed queries achieves 33% and 10% mAP for closed-set and open-set object localization, respectively, and significantly outperforms all related baselines.

Contributions of this paper are listed as follows:

1. We present an object proposal generation module that is guided by multimodal queries. It proposes a novel extension of our cross-modal attention scheme to generate a spatial compatibility matrix between the different query feature vectors and image features. Being query-aware, this module is capable of generating proposals even for those object categories that are unseen during training.
2. We propose a novel orthogonal-projection based proposal scoring scheme that can efficiently score queries from multiple modalities with the object proposals in a better way.
3. We demonstrate query-guided proposal generation and, finally, instance-level object localization on natural images using the query representation across modalities. Despite the large domain gap between the query (text and sketch) and the target (natural image) data points, we achieve impressive localization performance on challenging public benchmarks. Our method shows impressive performance gain ($\approx$ 4.7%) on open-set object localization.

The rest of the paper is organized as follows. Section 2 discusses the existing work from the computer vision literature that is related to this proposed research. Section 3 presents our proposed cross-modal object localization framework, including a detailed analysis of the novel orthogonal projection-based proposal scoring and the cross-modal attention scheme involving multiple query modalities. Section 4 demonstrates the effectiveness of our approach via performing extensive experiments using several publicly available datasets for various modalities, followed by a conclusion in Section 5.

2 Related work

2.1 Sketch for vision tasks

A better understanding of hand-drawn sketches and their utility to computer vision and cognitive science at large has been an active area of research. In order to achieve this goal, developing techniques for a robust representation of sketches has gained huge attention over the last decade. In addition to convolutional neural networks [85], which are traditionally used, there have been some works that utilize RNN [21] and transformers [81] for learning sketch encoders.

The area that has significantly benefited from sketch representation techniques is sketch-based image retrieval or SBIR. The goal of SBIR is to retrieve natural images using sketches as queries. Traditional SBIR methods utilize a separate feature computation step that uses manually-tuned features, such as SIFT or histogram of gradients, followed by a bag-of-words encoding as sketch representation [14, 26] and sometimes image edges or contours are also extracted for building image features [76, 90]. On the other hand, modern methods leverage deep networks for learning a joint embedding space where sketches and natural images are projected. In these works, often ranking loss such as the contrastive [64] or the triplet loss [86] is used to learn a ranking function between the sketch queries and the candidate images. In [70], researchers have leveraged an attention model to solve fine-grained SBIR and have also introduced higher-order learnable energy function-based loss to alleviate the domain gap between the images and the sketches. In [4, 10], researchers have tackled the task of noise-tolerant image retrieval. To improve the efficiency for large-scale image retrieval, hashing models have been explored [41, 66, 80, 91].

Sketches have also been used to study the perceptual grouping ability of machines [39, 54] and sketch synthesis [17, 21, 69]. In our earlier work [71], we have shown the utility of hand-drawn sketches for object localization in natural images. Although sketches provide critical visual cues, they often lack semantics. To fill this gap, in this work, we propose a method to leverage semantics (using object category name or gloss) along with sketches for object localization.

2.2 Visual grounding

Visual grounding [34, 42, 52, 77] is a task that has some similarities with the task presented in this paper. However, there are two key differences: (i) visual grounding often restricts itself to natural language query alone, whereas our model supports sketch, object category, and gloss as queries. (ii) The natural language query in visual grounding describes the object, its attributes, and its relationships with other objects in the image, and it is not an object definition (or gloss) like ours. Further, unlike visual grounding, which leverages large-scale

image-caption pairs during training, we only have very few unique definitions (or gloss) for every object and a large number of hand-drawn sketches for training our object localization framework.

2.3 Object detection

Object detection is a core computer vision task. Modern object detection methods can be grouped into the following two categories: (i) proposal-free methods [31, 37, 40, 58, 59, 65, 84] and (ii) proposal-based methods [5, 18, 19, 22, 23, 61, 92]. Proposal-free methods are single-stage detectors, and therefore, they are faster during inference. However, they often fall short of performance as compared to proposal-based approaches.

Under proposal-based approaches, Girshick et al. [19] have proposed a two-stage object detection method. In their first stage, they leverage selective search [72] to generate object proposals. In the second stage, these generated proposals were classified as one of the object categories using an independently-trained classifier. Ren et al. [61] proposed an end-to-end trainable object detector popularly known as Faster R-CNN. These object detectors are reasonably successful in the closed-set setting. However, they do not generalize well in an open set setting where an object category may or may not be seen during the training phase. Recently, Hsieh et al. [25] have proposed one-shot object detection. In their work, an object image is used as a query, and all the instances of the query object in the target image are detected. However, unlike their work, where query and target images are from the same distribution, i.e., natural images, our queries, i.e., gloss or hand-drawn sketches, are from a significantly different domain than those of the target images.

Object detection in the zero-shot setting has also been studied in the literature [3, 55, 56]. Typically by alleviating the confusion between the "background" and unseen class, these methods improve object proposal generation for unseen object categories [3]. Recently, there has been extensive research in context-aware zero-shot detection [8, 27, 43, 83] which incorporates joint detection of multiple objects [8, 27] or a background scene graph as a knowledge source [43]. These works are similar in spirit to the proposed work, but the proposed work is query-guided and utilizes a multimodal query to perform object localization.

2.4 Attention schemes in deep learning literature

The use of attention models is prevalent in deep learning literature. They allow the relevant features to become more crucial. Here, we briefly review the utilization of attention in object localization literature. Choe et al. [9] presented an attention network to score object proposals and showed its utility in object localization. Li et al. [35] proposed Attention to context Convolution Neural Networks (AC-CNN) in object detection to integrate local and global context. Leveraging the self-attention mechanism [78], Heish et al. [25] presented Co-attention and co-excitation network (CoAtEx) for one-shot object localization. In their work, the response at each feature map location of an image is computed as a weighted combination of the feature vectors at each feature map location of the query. Here weights depend on the similarity between target and query image pixel pairs. It should be noted that both query and target images are from the same modality in CoAtEx. In comparison, proposed cross-modal attention determines the spatial compatibility between global query representation and localized image region representations, thereby mitigating the domain misalignment. In more recent

work, authors [82] used class-specific attentive vectors inferred from images of objects in a meta-set to apply channel-wise soft attention to proposals' feature maps. The channel-wise soft attention may not be trivially utilized in our problem setup due to the domain gap between the target and the query.

2.5 Multimodal learning for vision

The natural environment of any visual task contains multiple modalities. Leveraging data from multiple modalities has been an expanding area of research in the vision literature. Multimodal learning has a very broad range of applications in computer vision, including but not limited to medical image analysis [29], audio-visual speech recognition [53], multimedia event detection [33], multimodal emotion recognition [68] and visual question answering [49]. A key challenge in this area is to summarize information from multiple modalities in a way that is lossless and exploits their complementary or supplementary nature. In [6, 20], the authors study the problem of emotion recognition by utilizing facial expressions, head gestures, and other visual cues. Researchers in [57, 62, 74, 88] utilize multi-view LSTM to model cross-view interactions over time or structured data. In the area of image retrieval, composing multimodal queries has gained interest in the last few years [75]. In [11], researchers incorporate semantic and geographical information to improve image retrieval, while researchers in [7] utilize both textual and visual features for improved image retrieval performance in the medical domain. To fuse multimodal input in information retrieval, concatenation [47, 50, 51] of features and multi-layer probabilistic latent semantic analysis (PLSA) [24] models [7] have been proposed. Attribute as operator [45], and parameter hashing [48] methods create a transformation matrix from text and use it to transform the image features. Researchers in [44, 63, 89] utilized visual cues for sentiment analysis in product and movie reviews, which conventionally used only text. They directly concatenated visual and textual representations in order to obtain a joint representation. Tensor fusion network [87] was proposed to fuse up to three different modalities for a multimodal sentiment analysis task. More recently, authors in [46] proposed attention bottlenecks in transformers to effectively fuse features from videos and audio, and in [1] three separate transformer models were trained using self-supervised learning with multi-modal contrastive losses to extract effective multi-modal representation from raw inputs of video, audio, and text. The reader is encouraged to read elaborate surveys [2, 73] on multimodal learning to know more about this area. Our work is closely related to this line of study, where we use cues from a natural image, a sketch, and text to perform object localization.

3 Our approach

In this section, we first provide a formal introduction to the multimodal query-guided object localization problem. Then, we present our solution by first describing the cross-modal attention scheme for either of the query modalities. We consider the specific examples of text and hand-drawn sketches as the query modalities in this paper and show how our cross-modal attention can be leveraged for each modality to generate proposals relevant to the query objects. We then introduce our novel orthogonal projection-based proposal scoring scheme to better score object proposals with respect to queries of multiple modalities.

3.1 Problem formulation

Let $I = I_{train} \bigcup I_{test}$ be a set of all-natural scene images in a dataset $\mathcal{D}_I$, each containing a variable number of object instances and categories. Here I_{train} and I_{test} are sets of train and test images, respectively. Like any other machine learning task, these two sets are mutually exclusive, and only I_{train} is available during training. Further, let $S = S_{train} \bigcup S_{test}$ be a set of all sketches, each containing one object, $T = T_{train} \bigcup T_{test}$ be the set of a textual description of an object (either object category name or gloss of the object category), and $C = C_{train} \bigcup C_{test}$ be a set of all object categories. During training, each training sample contains an image $i \in I_{train}$, a sketch query $s_c \in S_{train}$, a text query $t_c \in T_{train}$ where $c \in C_{train}$, and all the bounding boxes corresponding to object category c in the image i. At test time, given an image $i^{'} \in I_{test}$ and a sketch query $s_{c'} \in S_{test}$, a text query $t_{c'} \in T_{test}$, where $c^{'} \in C_{test}$, the problem is to localize all the instances of the object category $c^{'}$ in the image $i^{'}$. Note that we show experimental results in cases where $C_{train} = C_{test}$, i.e., categories in C_{test} are seen during training time (closed-set object localization), as well as $C_{train} \bigcap C_{test} = \phi$, i.e., categories in C_{test} are not seen during training (open-set object localization).

The proposed multimodal query-guided object localization is end-to-end trainable, and it works in the following two stages: (i) query-guided object proposal generation (Section 3.2), and (ii) orthogonal-projection based proposal scoring (Section 3.3). Figure 2 shows a schematic diagram of the proposed framework.

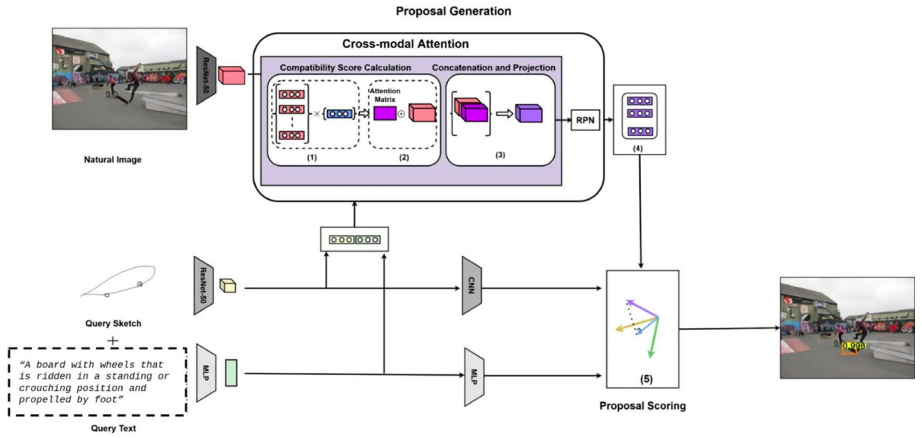

Fig. 2 Given an image and queries of different modalities, our object localization framework works in the following two stages: (i) **query-guided proposal generation:** in this step, the global fused feature vector of different queries that are shown using blue color is scored with the image feature vectors that corresponds to each location on the image feature map that is shown using pink color to generate the spatial compatibility also called the attention scores. (Block 1). Next, these attention scores, which are shown using violet color, are multiplied with the image feature maps, which are shown using pink color to get the attention features (Block 2). Before passing it through the region proposal network (RPN), it is first concatenated with the original feature maps and projected to the original dimension. The RPN is able to generate relevant object proposals because of the spatial compatibility, that is integrated into the image feature maps, between global fused queries representation and regional image representation (Block-3), (ii) **orthogonal-projection based proposal scoring:** the representation for each of the pooled object proposals that are shown using indigo is scored with query feature vectors from multiple modalities to generate localization for the object of interest (Block-5). The proposal vector is projected onto the subspace spanned by the queries, and the projection vector is utilized to query against the proposal vector. **[Best viewed in color].**

3.2 Cross-modal attention for query-guided object proposal generation

Faster R-CNN [61] is a popular framework for two-stage object detection, and in the first stage, it uses a region proposal network (RPN) to generate object proposals. A vanilla region proposal network could be used to generate object proposals in our task. However, a traditional RPN is not built to take advantage of any query-level information on object appearance or semantics. As a result, the object proposals that are relevant to the sketch or textual queries may not even be generated, particularly when the object of interest is of low resolution, occluded, hidden among other objects that are better represented in the input images, or most importantly, is one of the categories which is unseen during training. Therefore, using an RPN in its vanilla form may not suffice in our pipeline. To address the aforementioned problem, we proposed cross-modal attention to incorporate the sketch query information in the RPN in our earlier work [71]. In this work, we adapt the cross-modal attention to incorporate multimodal queries in the RPN and thereby guide the proposal generation. Regions of interest (ROIs) are pooled from region proposals generated using RPN utilizing a strategy similar to the Faster R-CNN, and a scoring function Θ is learned between these ROIs and joint representations of sketch and text queries.

We now describe our cross-modal attention framework to generate object proposals relevant to the queries of different modalities. A preliminary version of this, specific to sketch queries only, was presented in our earlier work. In this work, we extend the framework to include additional modalities, such as text queries. We feed a joint representation of sketch and text modalities to the proposal generation module, which is trained to produce a spatial weight map that provides high scores to the areas on the target image which are visually or semantically similar to the object corresponding to the given query(ies).

As mentioned earlier in this paper, we consider the examples of two query modalities, i.e., sketch and text. Suppose a sketch $s_c \in S$ and a text $t_c \in T$ (either category name or gloss) of an object category $c \in C$ is used to query an image $i \in I$. To generate the feature representation of images and sketches, we use ResNet-50 models pretrained on Imagenet [12] and Quick Draw [30] datasets, respectively, as backbones. We use either of the two types of text queries: object category name and generic object description (aka gloss). Feature representations for these are obtained using a language encoding scheme followed by a trainable multi-layer neural network. The de facto choice for language encoding now is fine-tuned BERT [13] model. We used them to represent the object category name and its gloss, respectively. Suppose ϕ_I, ϕ_T, and ϕ_S represent these backbone feature encoders, then image, text, and sketch feature maps are computed as:

$$i^{\phi_I} = \phi_I(i) \quad , \quad s_c^{\phi_S} = \phi_S(s_c) \quad \text{and} \quad t_c^{\phi_T} = \phi_T(t_c), \tag{1}$$

where, $i^{\phi_I} \in \mathbb{R}^{w \times h \times d}$, $t_c^{\phi_T} \in \mathbb{R}^d$, and $s_c^{\phi_S} \in \mathbb{R}^{w' \times h' \times d}$ are the extracted image, text, and sketch feature representations respectively. From these feature maps, the compatibility score is learned between the sketch and the text queries, and the image feature maps by first applying non-linear transformations as below:

$$i^{\psi_I} = \psi_I(i^{\phi_I}) \ , \ s_c^{\psi_S} = \psi_S(s_c^{\phi_S}) \quad \text{and} \quad t_c^{\psi_T} = \psi_T(t_c^{\phi_T}). \tag{2}$$

A set of local feature vectors is formed by obtaining one vector at each location (m, n) in the image feature map i^{ψ_I}, where $m \in \{1, 2, \ldots, w\}$ and $n \in \{1, 2, \ldots, h\}$. Each vector represents a spatial region on the target image, and the set gives us the spatial distribution of the features. Subsequently, this is compared against a fusion of global representation of the sketch features and textual features. For image feature map i.e. $i^{\psi_I} \in \mathbb{R}^{w \times h \times d}$, the extracted

set of feature vectors is represented as $L^i = \{\mathbf{L}^i_1, \mathbf{L}^i_2, ..., \mathbf{L}^i_{w\times h}\}$ where $\mathbf{L}^i_j \in \mathbb{R}^{1\times 1\times d}$ $\forall$ $j \in \{1, 2, \ldots, w \times h\}$.

In the case of sketches, a global representation of sketch feature maps is obtained via the global max pool ($\mathcal{GMP}$) operation, i.e., $\mathbf{L}^{s_c}_g = \mathcal{GMP}(s^{\psi_S}_c)$, where, $\mathbf{L}^{s_c}_g \in \mathbb{R}^{1\times 1\times d}$. The sketch and text representations are first passed through a linear layer, concatenated, and projected to obtain the final query representation.

$$\mathbf{L}^{q_c}_g = W[W_s(\mathbf{L}^{s_c}_g)^T; W_t t^{\psi_T}_c], \tag{3}$$

where, $W \in \mathbb{R}^{d\times 2d}$, is a projection matrix that maps the concatenated global sketch representation and text representation from a 2d-dimensional space to a d-dimensional space. The end-to-end training process then results in a high-quality fused representation that captures information from both modalities. A spatial compatibility score between $\mathbf{L}^i_j \in L^i$ and $\mathbf{L}^{q_c}_g$, is computed as follows:

$$\lambda\left(\mathbf{L}^i_j, \mathbf{L}^{q_c}_g\right) = \frac{\mathbf{L}^i_j \cdot \mathbf{L}^{q_c}_g}{\mathcal{K}}, \tag{4}$$

where $\mathcal{K}$ is a constant. For simplicity of notation, we will refer to the left-hand side of (4) as λ_{jg} from here onwards.

It should be noted that these compatibility scores are generated as a spatial map, which can be understood as a 2D-map representing attention weights. Therefore, in order to obtain attended feature maps, we perform element-wise multiplication of these compatibility scores and the original image feature map at each spatial location, i.e.,

$$i^{a_I}_j = i^{\phi_I}_j \odot \lambda_{jg}, \forall j \in \{1, 2, \cdots, w \times h\}. \tag{5}$$

This attention feature map aims to capture information about the location of objects in an image that shares high compatibility score with both the sketch query and the text query. Therefore, to incorporate this information, attention feature maps are concatenated along the depth with the original feature maps, i.e., $i^{\phi_I}_f = [(i^{a_I})^T; (i^{\phi_I})^T]^T$, where $i^{\phi_I}_f \in \mathbb{R}^{w\times h\times 2d}$. These concatenated feature maps are projected to a lower-dimensional space to obtain the final feature maps, which are subsequently passed through the RPN to generate object proposals relevant to the sketch query.

3.3 Orthogonal-projection based proposal scoring

Once a small set of proposals represented as R_i for $i \in I$ are pooled from all query-guided region proposals generated by the RPN, feature vectors for these proposals are computed along with the final feature vectors for sketch and text query, respectively, as follows.

$$r^{\phi'_I}_k = \phi'_I(r^{\phi_I}_k)\ ,\ s^{\phi'_S}_c = \phi'_S(s^{\phi_S}_c) \text{ and } t^{\phi'_T}_c = \phi'_T(t^{\phi_T}_c), \tag{6}$$

where $r^{\phi'_I}_k \in \mathbb{R}^d$ is generated using standard Faster R-CNN protocols, $s^{\phi'_S}_c \in \mathbb{R}^d$, $t^{\phi'_T}_c \in \mathbb{R}^d$, $r_k \in R_i$. The ϕ'_I and ϕ'_S are two separate multi-layer CNN followed by mean pool, and ϕ'_T is the multi-layer feed-forward neural network. In order to rank these object proposals with respect to the multimodal query representation, a scoring function Θ is learned. During training, the region proposals are labeled as foreground (or 1) when they have ≥ 0.5 intersection over union (IoU) with the ground truth bounding boxes and the objects in the bounding boxes belong to the same class as the query, or else they are labeled as background (or 0). Then, we minimize a margin rank loss between the representations of the generated object proposals

and the queries such that object proposals that contain the object of the same class as the queries are ranked higher.

Object proposals belong to a domain different from the queries, which themselves are from entirely different domains. Further, queries from different domains may capture different kinds of information, e.g., the sketch of an object captures the shape information, while on the other hand, text captures the semantics of the object. Therefore, we need to compare the object proposals against both these kinds of information to obtain a better score. In order to ensure better scoring, we propose orthogonal-projection-based proposal scoring. We begin by finding the proposal feature vector's projection in the subspace defined by the queries. We then use that projection to compute a score with the representation of the proposal. By being an orthogonal projection, we use the closest vector containing the complementary information present in the sketch and the text, in the query space, to the representation of the proposal for proposal scoring. We now describe our proposed orthogonal projection scheme in detail.

3.4 Orthogonal projection

We construct a vector subspace $\mathcal{M}$ by considering the queries $s_c^{\phi'_S}$ and $t_c^{\phi'_T}$ as the basis vectors. Then, we perform an orthogonal projection of the object proposal vectors into this subspace. This projection yields a vector that contains the complementary information present in $s_c^{\phi'_S}$ and $t_c^{\phi'_T}$ and is closest to the proposal vector. To obtain the orthogonal projection, we first define a matrix $B_c = [s_c^{\phi'_S}, t_c^{\phi'_T}] \in \mathbb{R}^{d\times 2}$, and the projection matrix is defined in terms of B_c as follows:

$$P_{R(\mathcal{M})} = B_c(B_c^T B_c)^{-1} B_c^T, \tag{7}$$

where $P_{R(\mathcal{M})}$ is the projection matrix on the range space of $\mathcal{M}$ i.e. $R(\mathcal{M})$. In order to obtain the fusion, we project r_k onto the $R(\mathcal{M})$, i.e.

$$q_k^c = P_{R(\mathcal{M})} r_k^{\phi'_I}, \tag{8}$$

where $q_k^c \in \mathbb{R}^d$ is the fused sketch and text feature vector corresponding to the object proposal r_k.

In order to learn the scoring function Θ, the object proposal feature vectors are concatenated with the feature vector obtained before. These concatenated feature vectors are passed through the scoring function (a one-layer neural network in our framework), and it predicts the foreground probabilities of the proposals with respect to the fused query. Let a_k be the predicted foreground probability for proposal $r_k \in R_i$, and it is given by $a_k = \Theta([(r_k^{\phi'_I})^T; (q_k^c)^T]^T)$,where, both $r_k^{\phi'_I}$ and q_k^c are defined in Section 3.3. Now, towards training the scoring function Θ, a label $y_k = 1$ or 0 is assigned to r_k depending on its overlap with a ground truth object bounding box, as defined in the previous paragraph. Instead of using a neural network, cosine similarity can also be used to compute the score. Motivated from [25], the loss function used in training is defined as:

$$L(R_i, s_c) = \sum_k \left\{ y_k \max(m^+ - a_k, 0) + (1 - y_k)\max(a_k - m^-, 0) + L_{MR}^k \right\} \tag{9}$$

$$L_{MR}^k = \sum_{l=k+1} \left\{ \mathbf{1}_{[y_l = y_k]} \max(\mathsf{a\_k} - \mathsf{a\_l} - m^-, 0) + \mathbf{1}_{[y_l \neq y_k]} \max(m^+ - \mathsf{a\_k} - \mathsf{a\_l}, 0) \right\}, \tag{10}$$

where m^+ and m^- are positive and negative margins, respectively. The above loss function consists of two parts: (i) In (9), the first part of the loss function assures that the object proposals that are overlapping with the ground truth object locations are predicted as foreground with high probability. (ii) The second part of the loss function, i.e., (10), is a margin-ranking loss that takes pairs of the proposals as input. It aids in reinforcing a greater division between prediction probabilities of foreground and background object proposals, and therefore, it improves the ranking of all the foreground proposals overlapping with the true location(s) of the object of interest. Both parts of this loss function in (9) are equally weighted during training. Additionally, a cross-entropy loss on the labeled (background or foreground) feature vectors of the region proposals and a regression loss on the predicted bounding box location deltas (same regression loss as in Faster-RCNN) with respect to the ground truth bounding box are used for training.

Moreover, using the orthogonal projection scheme described before can also be viewed as a fusion technique that can fuse a number of queries of multiple modalities without requiring any additional parameters. An important objective of a fusion technique is that the resultant fused representation has better utility than the individual queries. This property can be meaningfully encoded as the following equations:

$$d(r_k^{\phi'_I}, f(s_c^{\phi'_S}, t_c^{\phi'_T})) \leq d(r_k^{\phi'_I}, s_c^{\phi'_S}) \tag{11}$$

$$d(r_k^{\phi'_I}, f(s_c^{\phi'_S}, t_c^{\phi'_T})) \leq d(r_k^{\phi'_I}, t_c^{\phi'_T}), \tag{12}$$

where $d(\cdot, \cdot)$ is a suitable distance function and $f(\cdot, \cdot)$ is a function that fuses $s_c^{\phi'_S}$ and $t_c^{\phi'_T}$. The objective of enforcing this constraint is that by design, the representations of the query modalities must improve in utility on fusion, i.e., the fused representation should be closer to the feature obtained from an object proposal than any individual query features as measured by a suitable distance function. Utilizing the Orthogonal Projection Scoring (OPS) for scoring inherently enforces these constraints. The OPS scheme involves determining the orthogonal projection of the region proposal vector onto the subspace defined by the query vectors. Since the orthogonal projection of a vector on a subspace leads to the closest vector in that subspace, the proposed sceheme gives a vector that is closer or exhibit a smaller distance to the region proposal vector than either of the query vectors defining that subspace. This property of the orthogonal projections is also mathematically specified by (11) and (12). Therefore, it could be viewed as a fusion technique that utilizes the proposal representation for better scoring without requiring additional parameters.

4 Experiments and results

4.1 Datasets

We evaluate the performance of the proposed framework using the following datasets Fig. 3:

4.1.1 QuickDraw [30]

It is a large-scale hand-drawn sketch dataset. It contains 50 million hand-drawn sketches of 345 object categories in all. In our experiments, we selected those sketch categories that overlap with MS-COCO or PASCAL-VOC, as described in the subsequent paragraphs.

QuickDraw sketches are stored as vector graphics, and we rasterized the sketches before feeding them into the ResNet.

4.1.2 MS-COCO [38]

It is a de facto natural scene dataset for studying object detection. It contains object bounding box annotations for 80 object categories. Between MS-COCO and QuickDraw datasets, 56 object categories are common. Therefore, we randomly selected a total of $800K$ sketches across these common classes for our experiments. The model is trained on the COCO-Train-2017 and evaluated on the MS-COCO-Val-2017 dataset.

4.1.3 PASCAL VOC [15]

It is another common object detection dataset. It contains a total of 20 object classes. We choose images of nine object categories that are common to the QuickDraw dataset for our experiments. We trained our model on the union of VOC2007 train-val and VOC2012 train-val sets and evaluated on the VOC-test-2007 set.

4.1.4 Gloss dataset

Semantic information about the object is introduced to our localization framework by utilizing an embedding of a brief sentence describing an object category, also known as gloss. We collected a gloss of object categories selected from the Visual Genome [32] and MS-COCO datasets from WordNet [16]. We refer to this collection as the Gloss dataset. This dataset contains gloss for *1615* object categories in all. Some examples of this dataset include gloss for sofa is *an upholstered seat for more than one person*, gloss for carrot is *deep orange edible root of the cultivated carrot plant*.

4.2 Baselines and our variants

In order to demonstrate the superior performance of our approach, we adapt and compare it with the following popular approaches from the object detection and image-guided localization literature:

4.2.1 Sketch-only baselines

In this section, we describe the baselines to evaluate the sketch-only object localization.
Modified Faster R-CNN [61]: For query-guided object localization tasks with a sketch query, we adapt Faster RCNN. Towards this end, during training, if an object instance in an image belongs to the same class as the sketch query, we assign class label 1 to it and 0 otherwise. We then generate object proposals using the vanilla region proposal network (RPN). Each region proposal is then identified as background or foreground using a binary classifier. To this end, the region of interest features for each region proposal is first concatenated with the query features and then passed through the binary classifier. We also used a triplet loss to rank the object region proposals concerning the sketch query. This baseline is referred to as modified Faster R-CNN in this paper.
Co-attention and co-excitation network (CoATex) [25]: It is a recent one-shot object localization method using image queries. The query information is integrated into the image feature

maps by utilizing non-local neural networks [78] and channel co-excitation [28]. This method is adapted to work with the sketch query directly, and it is used as a second baseline in our experiments.

The feature extractors for images and sketches are ResNet-50 models pre-trained on Imagenet and QuickDraw, respectively, for both these baseline methods.

4.2.2 Our variants

In order to perform a comparative study with the above-mentioned baseline approaches, we present the following variants of our approach:
Sketch only [71]: In this variant of our approach, we only use sketch queries to localize objects in our framework.
Gloss only: In this variant of our approach, we only use natural language description (aka gloss) of object categories to localize them in a natural scene. This variant is useful to demonstrate the effectiveness of our approach in cases there is no visual query available.
Sketch+Gloss: This is our full model. In this, we fuse two modalities, namely visual (sketch) and textual (gloss), using the following different fusion strategies. (i) **Late fusion:** Let the sets R^s and R^t be the set of proposals obtained after comparing with s_c and t_c respectively at the test time, where s_c and t_c are sketches and text queries of class $c \in C$ respectively. In late fusion, we take the union of these sets of proposals and choose the Top-N proposals as the final set. (ii) **Concatenation fusion:** Let s_c and t_c be the sketch and text queries, respectively. In this fusion strategy, these queries are concatenated and projected to obtain the fused query. (iii) **Proposed OPS as fusion:** Finally, we use the proposed orthogonal projection scoring (OPS) presented in Section 3.3 to fuse sketch and gloss embeddings.

4.3 Evaluation metric

Given the similarities between query-guided object localization and object detection tasks, we have utilized the mean Average Precision (mAP) and Average Precision at an IoU threshold of 0.5 (AP@50) to evaluate the efficacy of the localization methods. Initially, the Intersection over Union (IoU) is used to determine the degree of overlap between the ground truth bounding box and the generated bounding box. If the IoU is above a certain threshold, the generated bounding box is considered a true positive detection. The mAP is then computed as the average of the AP scores at different IoU thresholds, where AP is the precision value averaged over all the recall values. When calculating the mAP, the IoU threshold typically ranges from 0.5 to 0.95, while AP@50 is calculated at a fixed IoU threshold of 0.5. AP@50 evaluates the models' performance at a specific IoU threshold of 0.5, which is widely adopted in the field of object detection. In contrast, mAP provides a comprehensive evaluation of the model's performance across a range of IoU thresholds.

4.4 Experimental setup

In order to get the feature representation for the images and the sketches, we used two ResNet-50 models pre-trained on Imagenet [12] and a subset of 5 million images from QuickDraw [30], respectively. We use hand-drawn sketches from the common classes of QuickDraw to localize objects in images from MS-COCO and PASCAL-VOC datasets. Once the gloss dataset is created, we use WordNet synset matching to retrieve a set of similar object categories for each class. This set of similar categories is utilized to fine-tune a pretrained

BERT model [13] under the objective that similar classes' representation is close to each other than the non-similar classes. We evaluate the performance of our model under closed-set and open-set settings.

4.4.1 Open-set experimental setting

In the open-set experimental setting, out of the 56 common classes across COCO and QuickDraw, we choose 42 and 14 classes as 'seen' and 'unseen' categories, respectively. The 'seen' and 'unseen' splits are mutually exclusive in terms of object categories and labeled bounding boxes present to ensure the one-shot open-set experimental setting. Our model is trained exclusively on the dataset from the 'seen' classes, and only the 'unseen' classes are used for open-set evaluation. Similarly, for the PASCAL-VOC, out of the nine classes common to QuickDraw, three and six are chosen arbitrarily as 'unseen' and 'seen' categories, respectively. The image encoder is pretrained on the Imagenet dataset, except for 14 'unseen' classes and all associated classes obtained by matching their WordNet synsets. Similarly, except for the 14 categories in the 'unseen' set, the sketch encoder is pretrained using all of the QuickDraw categories.

4.4.2 Closed-set experimental setting

In this experimental setting, all the 56 common classes in MS-COCO and Quickdraw datasets are used during training, and the model is evaluated on all 56 categories at test time. Similarly, for the PASCAL-VOC dataset, all data points which correspond to 12 classes, which are common with QuickDraw, are used during the training. During the evaluation, the dataset from all 12 classes is utilized.

4.5 Implementation details

We use PyTorch v1.0.1 with CUDA 10.0 and CUDNN v7.1 to train the model using stochastic gradient descent (SGD) with a momentum of 0.9 on one NVIDIA 1080-Ti with a batch size of 10. The learning rate was initially set at 0.01, but it decays with a rate of 0.1 after every four epochs, and it is trained for 30 epochs. The constant $\mathcal{K}$ in (4) is fixed at 256 and $m^+ = 0.3$ and $m^- = 0.7$ in (9) and (10) for all experiments. To obtain the sentence embeddings average of the features of the final layer of BERT is utilized. The BERT model is fine-tuned with a learning rate of $5e-5$, and triplet loss is used during fine-tuning along with the hard-negative mining on the mini-batch. For optimal results, the cross-modal attention model is trained incrementally. Firstly, the localization model is trained without attention. Then, the attention model is added to it, and it is trained again. The training protocol is the same as explained before, and it is the same for both steps.

4.6 Results and discussion

We now quantitatively and qualitatively evaluate our model in different settings on the MS-COCO and PASCAL-VOC datasets. Our model on the MS-COCO dataset is compared against other related approaches in Table 1 in both open and closed-set settings. For the sketch-only experiments, the proposed cross-modal attention model significantly outperforms both modified Faster-RCNN and CoATex-based baselines. This is primarily because, unlike faster

Table 1 *Results in one-shot open-set and closed-set settings on the MS-COCO-Val-2017 dataset*

Method	Fusion	Open Set		Closed Set	
		%AP@50	%mAP	%AP@50	%mAP
Modified Faster RCNN	–	7.4	5.4	31.5	18.0
CoATex [25]	–	12.4	6.3	48.5	28.0
Ours					
Sketch only [71]	–	15.0	7.4	50.0	30.1
Gloss only	–	15.2	7.6	54.2	32.7
Sketch + gloss	Late	16.0	7.8	53.3	32.5
Sketch + gloss	Concat	18.8	9.6	53.4	32.6
Sketch + gloss	OPS	**19.7**	**10.0**	**54.4**	**33.0**

Bringing semantics using the additional queries, such as gloss and object category names, generally has a positive effect on the localization performance. Further, orthogonal projection-based scoring clearly outperforms other fusion techniques in challenging open-set settings

R-CNN, the cross-modal attention framework effectively incorporates the query information using spatial compatibility (attention) maps to generate region proposals that are relevant. Further, the CoATex baseline [25] utilizes the non-local feature maps and channel co-excitation module, and these modules are sensitive toward the domain gap present between query and image feature maps in our task. The proposed method, on the other hand, addresses this by computing a spatial compatibility (attention) map directly. Our model, by virtue of cross-modal attention, integrates query information in the image feature map before feeding it through the region proposal network. Consequently, our model is intrinsically able to generate relevant object proposals even for unseen object categories. As a result, our approach outperforms the baselines on unseen object categories.

However, when comparing the sketch-only model with the gloss-only model, the gloss-only model performs significantly better. Both these models have similar architectures aside from the modality of the query. Therefore, it indicates that text-only queries contain information that can be utilized better for object localization. Furthermore, we combine both the sketch query and the gloss query in our method, utilizing orthogonal-projection based scoring, and find that it leads to significant improvement in the localization performance; for example, in the close-set setting, we get 3.8% improvement when sketch queries are used along with object gloss. This improvement is even more significant (i.e., 4.7%) in an open-set setting, indicating that incorporating semantic information and shape information helps create a better representation for object localization (Refer Table 1).

Compared with the multimodal fusion baselines, the proposed orthogonal-projection-based proposal scoring scheme shows significant performance improvement, indicating that the proposed scheme is able to better capture the complementary information present in multiple modalities. The late fusion technique combines the predicted localization from each

Table 2 *Results in one-shot open-set setting on VOC test-2007 dataset*

Method	mAP
Modified Faster RCNN	0.65
CoATex [25]	0.61
Ours (Sketch only) [71]	0.65

Table 3 *Results in one-shot open-set setting on MS-COCO-Val-2017 dataset.*

Modality	%AP@50	%mAP
Sketch only	15.0	7.4
Gloss only	15.2	7.6
Sketch + gloss	19.7	10.0
Sketch + gloss with class	20.3	10.6
Sketch + gloss + class	**23.6**	**12.9**

This table shows if the category name is available for query, the localization performance of our method can be further improved

of the queries, indicating that combining predictions from multiple sub-optimal queries does not give a sufficient improvement in performance. The concatenation fusion does not impose any constraints on the fusion output and, therefore, leads to sub-optimal results.

The results for the PASCAL-VOC dataset are reported in Table 2. PASCAL-VOC is a small-size dataset, and for our experimental setting, it contains a small number of training images ($\approx$ 9K) with inadequate variability between the classes present in the dataset. It should be noted here that the training set contains only nine classes that are common with the QuickDraw dataset. The proposed method is comparable to the modified faster-RCNN baseline, suggesting that query-guided object localization is hard in case of insufficient data. The CoATex baseline suffers degradation in performance, indicating that it is unable to integrate sketch information in the image feature map during proposal generation in the case of small data size and large domain gap. Due small training size, gloss query experiments are not performed on PASCAL-VOC.

4.6.1 Ablation study

Instead of the gloss of an object, the class name of an object can also be used as a query modality. We used class in two ways: i) append the object class at the beginning of the gloss and then use it as a query, and ii) use the word2vec embedding of the object class along

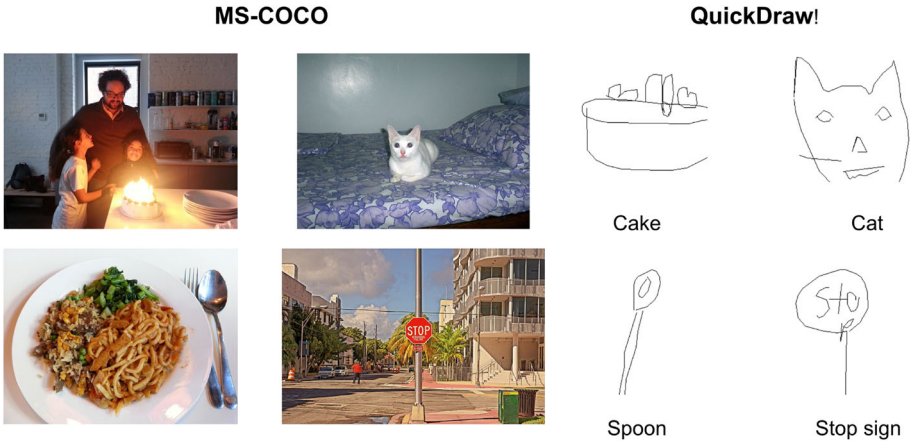

Fig. 3 We have shown some examples of the query sketches from the QuickDraw! and target images from the MS-COCO dataset

Table 4 Effect of m and $\mathcal{K}$ on the localization performance

	m			$\mathcal{K}$		
	0.35	0.3	0.25	200	256	312
mAP	9.8	10.0	8.2	9.1	10.0	9.5
%AP@50	19.5	19.7	16.2	18.3	19.7	19.0

The experiments are performed in *MS-COCO* dataset

with the object gloss and the sketch. In Table 3, gloss with class refers to the case when the object category is appended in front of the gloss. The results in Table 3 suggest that using the class name, if available, helps in performance improvement. Moreover, using the word2vec embeddings of the class name along with the sketch and the gloss gives a 3.9% improvement in performance. However, word2vec embeddings are trained on a large dataset set to learn semantic similarity between words, and therefore it violates the true open set experimental setting (Fig. 3).

4.6.2 Effect of *m* and $\mathcal{K}$

In this experiment, we studied the effect of margin m and scaling factor K on the localization performance of the model (Refer Table 4). The proposed model is fairly robust to changes in these parameters, and the best results are obtained when m = 0.3 and K = 256.

4.6.3 Comparison across different dataset splits

In this experiment, we compared our proposed Orthogonal Proposal Scoring (OPS) with the Concatenation fusion on different splits of train and test categories. As shown in Table 5, the performance of both of these methods varies across the splits, and our proposed OPS method performs the best across all splits.

4.6.4 Additional experiments

Sketches are heterogeneous in quality, and they often tend to capture complementary information on an object's shape, characteristics, and appearance, and many times multiple sketches of an object can be utilized. Therefore, we also compare the proposed multimodal localization with multiple sketch-based localization, and in order to utilize multiple sketches, we use the following two fusion techniques [71]: **(i) Feature Fusion** Image feature maps for different sketch queries are first generated, and then global max pool operation is applied to fuse these feature maps. Let an image be queried by N sketches that belong to the same object category $c \in C$, which is denoted as set $\{s_c^1, \ldots, s_c^N\}$. These queries are then fed through the sketch

Table 5 The performance comparison of the proposed OPS scoring with the Concatenation fusion for different sets of classes in the Open set

Fusion	Split1 AP@50	Split 2 AP@50	Split 3 AP@50	Split 4 AP@50	Avg AP@50
Concat	18.8	16.4	18.1	17.8	17.8
OPS	**19.7**	**17.6**	**19.9**	**18.9**	**19.0**

We used the sketch of the class along with the Gloss of the class in this experiment. The experiments are performed on *MS-COCO* dataset

backbone network, and suppose the representation learned for the n^{th} sketch is denoted as ${}^{n}s_c^{\phi_S}$. These feature map representations for each query are concatenated together to yield a composite feature map $R^{w \times h \times d \times N}$. Finally, a global max pool operation is performed across all N channels to obtain a fused feature map for the sketch queries.

(ii) Attention Fusion For each of the N sketch queries, attention maps are first generated and concatenated. Then depth-wise mean pool operation is applied to obtain the resultant fused attention map, which is used as input to the object localization pipeline (Section 3.3).

We perform an experiment where we use multiple (five) sketch queries instead of one, and these results are reported in Table 1. We observe that using multiple sketch queries improves the localization performance compared to using just a single sketch query. However, utilizing multiple modalities seems to perform better than fusing multiple sketch queries for both the open-set and closed-set experimental settings (refer to Table 6). It could be attributed to the fact that the textual data are pretrained such that it captures the semantic similarity of the object categories and, therefore, captures complementary information from the visual representation.

4.6.5 Qualitative results and failure analysis

To illustrate the effect of introducing the gloss as a query and the sketch, we visualize the localization results when the only sketch or gloss is available for the query and when both modalities are available. As shown in Fig. 4, the gloss of an object is able to assist the sketch query to generate better localization for the case of open-set queries. These visualizations, along with the empirical results, illustrate that using semantic information from the gloss and shape information from the sketch helps improve the localization performance for unseen categories. In the fifth row, the model is not able to discriminate even when both the sketch and the gloss are available for query. Similarly, in the sixth row, the model is confused about the object represented by the queries and is only able to localize the part of the object. Moreover, when localization for either of the query is correct, the combined model is able to localize the object with better confidence.

Moreover, in Fig. 5, we showed some qualitative results in which we query the same image with two different queries belonging to separate classes. As shown in the figure, the model is able to localize the correct objects.

Table 6 Comparison of the multi-sketch localization with multimodal localization in multi-query closed-set and open-set categories setting on the MS-COCO-val-2017 dataset

Method	Open Set		Closed Set	
	%AP@50	%mAP	%AP@50	%mAP
Ours (Sketch only) [71]	15.0	7.4	50.0	30.1
+Feature Fusion(3 Sketches)	17.1	7.3	51.9	31.0
+Feature Fusion (5 Sketches)	16.3	7.6	52.6	32.0
+Attention Fusion(3 Sketches)	17.6	7.5	52.0	31.0
+Attention Fusion (5 Sketches)	17.1	8.0	53.1	32.0
+Gloss	**19.7**	**10.0**	**54.4**	**33.0**

Here, five sketches mean we use five randomly selected sketch queries. Further, Open Set and Closed Set represent disjoint and common train and test categories, respectively

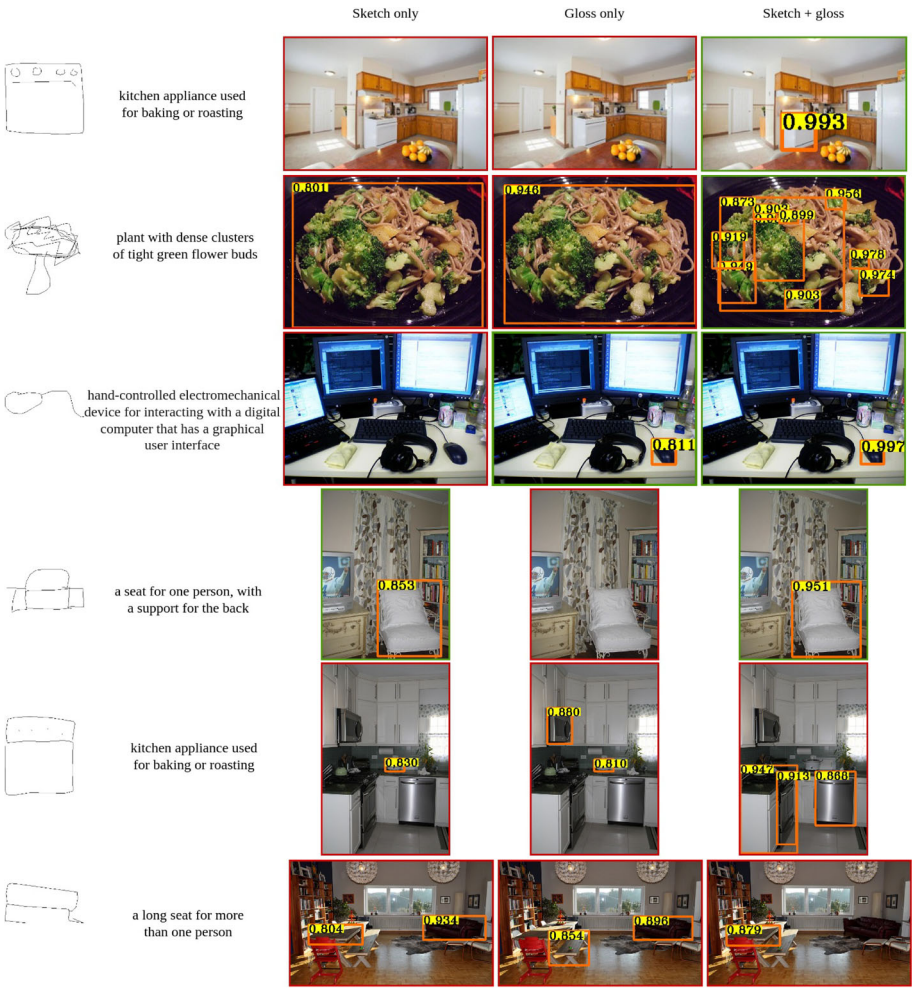

Fig. 4 The localization results are shown for the case when only the sketch query (third column), only gloss query (fourth column), and both sketch and gloss queries (fifth column). The results are shown for the open-set setting, i.e., these categories are unseen during training. The first two columns show sketch and gloss queries. We observe that having gloss brings semantics to the model and thereby enables it to perform better than *sketch only* localization. The last two rows show some of the failure cases

4.6.6 Comparison on computational time

Our model is computationally efficient. On average, it takes 0.08 seconds per query (sketch + gloss) to localize objects in the target scene. Compared to this CoATex [25] and modified faster RCNN takes 0.07 and 0.06 seconds per query, respectively, for localizing objects in the target scene (Refer Table 7). All these experiments are run on a system with Nvidia 1080-Ti GPU (with 11 GB VRAM) on Intel Xeon 4208 CPU (64 GB RAM).

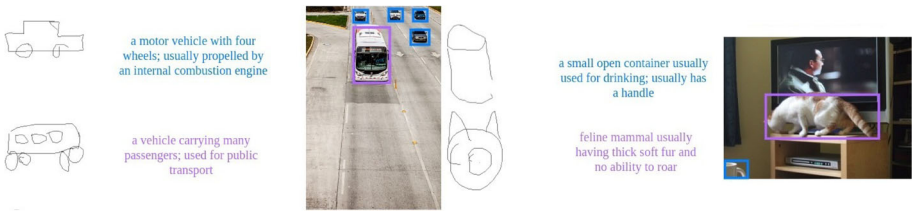

Fig. 5 The **multi-target localization results** are shown for the case when both sketch and gloss query is available. The results are shown for the open-set setting. The first column shows the sketch queries, and the second column shows the corresponding gloss queries using two different colors, i.e., blue and purple. Corresponding localizations in the target image are shown using the same colors as the gloss queries. **[Best viewed in color]**

Table 7 The average inference time comparison of the models

Model	Inference time (in seconds)
Modified Faster RCNN	0.06
CoATex [25]	0.07
Ours	0.08

All the models are evaluated on the same system with Nvidia 1080-Ti GPU (with 11 GB VRAM) on Intel Xeon 4208 CPU (64 GB RAM)

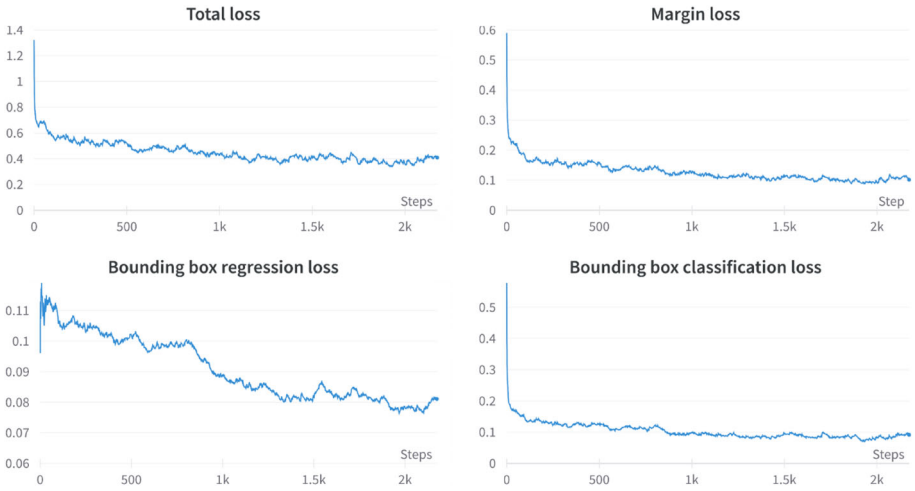

Fig. 6 The presented plots illustrate the progression of the training loss curves for the proposed model

4.6.7 Training loss curves

The Fig. 6 show the progression of the total training loss along with the component loss.

5 Conclusion

In this paper, we have investigated multimodal query-guided object localization in natural images. Our proposed framework seamlessly fuses sketch and text queries and generates object proposals that are relevant to the query. We further proposed a novel proposal scoring mechanism using the orthogonal projection. The noticeable performance gain achieved over the baselines establishes the efficacy of the proposed framework. Moreover, the proposed framework, by virtue of query-guided proposal generation and our novel proposal scoring scheme, is also effective for open-set object localization. We have performed extensive experiments and demonstrated the utility of bringing semantics using gloss in the object localization framework. Our work further strengthens the argument in the literature, i.e., effectively using information across multiple modalities and exploiting their complementary nature can improve performance on learning tasks.

Funding This work is partly supported by research grants from the Advanced Data Management Research Group, Corporate Technologies, Siemens Technology and Services Pvt. Ltd., and Pratiksha Trust, Bengaluru, India.

Data availibility statement The data that support the findings of this study are publicly available at MS–COCO [38], QuickDraw! [21], and PASCAL-VOC [15].

References

1. Akbari H, Yuan L, Qian R, Chuang W-H, Chang S-F, Cui Y, Gong B (2021) Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Proceedings of the Conference on Neural Information Processing Systems (NIPS) 34:24206–24221
2. Baltrušaitis T, Ahuja C, Morency L-P (2018) Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 41(2):423–443
3. Bansal A, Sikka K, Sharma G, Chellappa R, Divakaran A (2018) Zero-shot object detection. In: Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp 384–400
4. Bhunia AK, Koley S, Khilji AFUR, Sain A, Chowdhury PN, Xiang T, Song Y-Z (2022) Sketching without worrying: Noise-tolerant sketch-based image retrieval. ArXiv arXiv:2203.14817
5. Cai Z, Vasconcelos N (2018) Cascade R-CNN: Delving into high quality object detection. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp6154–6162
6. Calder AJ, Burton AM, Miller P, Young AW, Akamatsu S (2001) A principal component analysis of facial expressions. Vis Res 41(9):1179–1208
7. Cao Y, Steffey S, He J, Xiao D, Tao C, Chen P, Müller H (2014) Medical image retrieval: a multimodal approach. Cancer Informat 13:14053
8. Chen Z, Huang S, Tao D (2018) Context refinement for object detection. In: Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp 71–86
9. Choe J, Shim H (2019) Attention-based dropout layer for weakly supervised object localization. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 2219–2228
10. Chowdhury PN, Bhunia AK, Gajjala VR, Sain A, Xiang T, Song Y-Z (2022) Partially does it: Towards scene-level fg-sbir with partial input. ArXiv arXiv:2203.14804
11. Dang-Nguyen D-T, Boato G, Moschitti A, De Natale FG (2012) Supervised models for multimodal image retrieval based on visual, semantic and geographic information. In: Proceedings of the International Workshop on Content-Based Multimedia Indexing (CBMI), pp 1–5 IEEE
12. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 248–255

13. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 4171–4186
14. Eitz M, Hildebrand K, Boubekeur T, Alexa M (2010) Sketch-based image retrieval: Benchmark and bag-of-features descriptors. IEEE Trans Vis Comput Graph (TVCG) 17(11):1624–1636
15. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis (IJCV) 88(2):303–338
16. Fellbaum C (2012) Wordnet. The encyclopedia of applied linguistics
17. Ge S, Goswami V, Zitnick CL, Parikh D (2020) Creative sketch generation. arXiv preprint arXiv:2011.10039
18. Girshick R (2015) Fast R-CNN. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 1440–1448
19. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 580–587
20. Glodek M, Tschechne S, Layher G, Schels M, Brosch T, Scherer S, Kächele M, Schmidt M, Neumann H, Palm G et al (2011) Multiple classifier systems for the classification of audio-visual emotional states. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII), pp 359–368 Springer
21. Ha D, Eck D (2017) A neural representation of sketch drawings. Proceedings of the International Conference on Learning Representations (ICLR)
22. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 37(9):1904–1916
23. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 2980–2988
24. Hofmann T (2013) Probabilistic latent semantic analysis. Proceedings of the conference on Uncertainty in Artificial Intelligence (UAI)
25. Hsieh T-I, Lo Y-C, Chen H-T, Liu T-L (2019) One-shot object detection with co-attention and co-excitation. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS), pp 2721–2730
26. Hu R, Collomosse J (2013) A performance evaluation of gradient field hog descriptor for sketch based image retrieval. Comp Vision Image Underst (CVIU) 117(7):790–806
27. Hu H, Gu J, Zhang Z, Dai J, Wei Y (2018) Relation networks for object detection. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 3588–3597
28. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 2011–2023
29. James AP, Dasarathy BV (2014) Medical image fusion: A survey of the state of the art. Information Fusion 19:4–19
30. Jongejan J, Rowley H, Kawashima T, Kim J, Fox-Gieg N (2016) The quick, draw!-ai experiment. https://quickdraw.withgoogle.com
31. Kong T, Sun F, Liu H, Jiang Y, Shi J (2019) Foveabox: Beyond anchor-based object detector. IEEE Trans Image Process (TIP), 7389–7398
32. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma DA et al (2017) Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis (IJCV) 123(1):32–73
33. Lan Z-z, Bao L, Yu S-I, Liu W, Hauptmann AG (2014) Multimedia classification and event detection using double fusion. Multimedia Tools and Applications 71(1):333–347
34. Li X, Jiang S (2018) Bundled object context for referring expressions. IEEE Transactions on Multimedia (TOMM) 20(10):2749–2760
35. Li J, Wei Y, Liang X, Dong J, Xu T, Feng J, Yan S (2017) Attentive contexts for object detection. IEEE Transactions on Multimedia (TOMM) 19(5):944–954
36. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit, pp 2117–2125
37. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 2999–3007
38. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft COCO: Common objects in context. In: Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp 740–755
39. Li K, Pang K, Song J, Song Y-Z, Xiang T, Hospedales TM, Zhang H (2018) Universal sketch perceptual grouping. In: Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp 582–597

40. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: Single shot multibox detector. In: Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp 21–37
41. Liu L, Shen F, Shen Y, Liu X, Shao L (2017) Deep sketch hashing: Fast free-hand sketch-based image retrieval. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 2298–2307
42. Liu Y, Wan B, Zhu X, He X (2020) Learning cross-modal context graph for visual grounding. In: Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), pp 11645–11652
43. Luo R, Zhang N, Han B, Yang L (2020) Context-aware zero-shot recognition. In: Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), pp 11709–11716
44. Morency L-P, Mihalcea R, Doshi P (2011) Towards multimodal sentiment analysis: Harvesting opinions from the web. In: Proceedings of the International Conference on Multimodal Interfaces (ICMI), pp 169–176
45. Nagarajan T, Grauman K (2018) Attributes as operators: factorizing unseen attribute-object compositions. In: Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp 169–185
46. Nagrani A, Yang S, Arnab A, Jansen A, Schmid C, Sun C (2021) Attention bottlenecks for multimodal fusion. Proceedings of the Conference on Neural Information Processing Systems (NIPS) 34:14200–14213
47. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011) Multimodal deep learning. In: Getoor L, Scheffer T (eds.) Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011. Omnipress, ??? pp 689–696 https://icml.cc/2011/papers/399_icmlpaper.pdf
48. Noh H, Hongsuck Seo P, Han B (2016) Image question answering using convolutional neural network with dynamic parameter prediction. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit, pp 30–38
49. Osman A, Samek W (2019) Drau: Dual recurrent attention units for visual question answering. Comp Vision Image Underst (CVIU) 185:24–30
50. Pérez-Rosas V, Mihalcea R, Morency L (2013) Utterance-level multimodal sentiment analysis. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers. The Association for Computer Linguistics, ??? pp 973–982 https://aclanthology.org/P13-1096/
51. Pham T-T, Maillot NE, Lim J-H, Chevallet J-P (2007) Latent semantic fusion model for image retrieval and annotation. In: Proceedings of the ACM Conference on Conference on Information and Knowledge Management (CIKM), pp 439–444
52. Plummer BA, Kordas P, Hadi Kiapour M, Zheng S, Piramuthu R, Lazebnik S (2018) Conditional image-text embedding networks. In: Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp 249–264
53. Potamianos G, Neti C, Gravier G, Garg A, Senior AW (2003) Recent advances in the automatic recognition of audiovisual speech. Proc IEEE 91(9):1306–1326
54. Qi Y, Song Y-Z, Xiang T, Zhang H, Hospedales T, Li Y, Guo J (2015) Making better use of edges via perceptual grouping. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 1856–1865
55. Rahman S, Khan S, Barnes N (2018) Polarity loss for zero-shot object detection. arXiv preprint arXiv:1811.08982
56. Rahman S, Khan S, Porikli F (2018) Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In: Proceedings of the Asian Conference on Computer Vision (ACCV), pp 547–563 Springer
57. Rajagopalan SS, Morency L-P, Baltrusaitis T, Goecke R (2016) Extending long short-term memory for multi-view structured learning. In: Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp 338–353 Springer
58. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 779–788
59. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767
60. Ren S, He K, Girshick R, Sun J (2016) Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 39(6):1137–1149
61. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS), pp 91–99
62. Ren J, Hu Y, Tai Y-W, Wang C, Xu L, Sun W, Yan Q (2016) Look, listen and learn-a multimodal lstm for speaker identification. In: Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI), vol 30

63. Rosas VP, Mihalcea R, Morency L-P (2013) Multimodal sentiment analysis of spanish online videos. IEEE Intell Syst 28(3):38–45
64. Sangkloy P, Burnell N, Ham C, Hays J (2016) The sketchy database: learning to retrieve badly drawn bunnies. ACM Trans Graph (TOG) 35(4):1–12
65. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, Lecun Y (2013) Overfeat: Integrated recognition, localization and detection using convolutional networks. Proceedings of the International Conference on Learning Representations (ICLR)
66. Shen Y, Liu L, Shen F, Shao L (2018) Zero-shot sketch-image hashing. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 3598–3607
67. Sivic J, Zisserman A (2003) Video Google: A text retrieval approach to object matching in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 1470–1477
68. Soleymani M, Pantic M, Pun T (2011) Multimodal emotion recognition in response to videos. IEEE Trans Affect Comput (TAC) 3(2):211–223
69. Song Y-Z (2020) Bézier sketch: A generative model for scalable vector sketches. Proceedings of the proceedings of the European Conference on Computer Vision (ECCV) 2020:632–647
70. Song J, Yu Q, Song Y-Z, Xiang T, Hospedales T.M (2017) Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 5552–5561
71. Tripathi A, Dani RR, Mishra A, Chakraborty A (2020) Sketch-guided object localization in natural images. In: Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp 532–547 Springer
72. Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW (2013) Selective search for object recognition. International Journal of Computer Vision (IJCV) 104(2):154–171
73. Uppal S, Bhagat S, Hazarika D, Majumder N, Poria S, Zimmermann R, Zadeh A (2022) Multimodal research in vision and language: A review of current and emerging trends. Information Fusion 77:149–171
74. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: A neural image caption generator. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 3156–3164
75. Vo N, Jiang L, Sun C, Murphy K, Li L, Fei-Fei L, Hays J (2019) Composing text and image for image retrieval - an empirical odyssey. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 6439–6448
76. Wang S, Zhang J, Han TX, Miao Z (2015) Sketch-based image retrieval through hypothesis-driven object boundary selection with hlr descriptor. IEEE Transactions on Multimedia (TOMM) 17(7):1045–1057
77. Wang L, Li Y, Huang J, Lazebnik S (2018) Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 41(2):394–407
78. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 7794–7803
79. Wang C, Ren W, Huang K, Tan T (2014) Weakly supervised object localization with latent category learning. In: Fleet DJ, Pajdla T, Schiele B, Tuytelaars T (eds.) Proceedings of the Proceedings of the European Conference on Computer Vision, vol. 8694. Springer, ??? pp 431–445
80. Xu P, Huang Y, Yuan T, Pang K, Song Y-Z, Xiang T, Hospedales TM, Ma Z, Guo J (2018) Sketchmate: Deep hashing for million-scale human sketch retrieval. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 8090–8098
81. Xu P, Joshi CK, Bresson X (2021) Multigraph transformer for free-hand sketch recognition. IEEE Transactions on Neural Networks and Learning Systems, 1–12
82. Yan X, Chen Z, Xu A, Wang X, Liang X, Lin L (2019) Meta R-CNN: Towards general solver for instance-level low-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 9576–9585
83. Yang J, Lu J, Lee S, Batra D, Parikh D (2018) Graph r-cnn for scene graph generation. In: Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp 670–685
84. Yi J, Wu P, Metaxas DN (2019) ASSD: attentive single shot multibox detector. Comp Vision Image Underst (CVIU) 189
85. Yu Q, Yang Y, Liu F, Song Y-Z, Xiang T, Hospedales TM (2017) Sketch-a-net: A deep neural network that beats humans. Int J Comput Vis (IJCV) 122(3):411–425
86. Yu Q, Liu F, Song Y-Z, Xiang T, Hospedales TM, Loy C-C (2016) Sketch me that shoe. In: Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR), pp 799–807
87. Zadeh A, Chen M, Poria S, Cambria E, Morency L-P (2017) Tensor fusion network for multimodal sentiment analysis. Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP), 1103–1114

88. Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency L-P (2018) Memory fusion network for multi-view sequential learning. In: Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI), vol 32
89. Zadeh A, Zellers R, Pincus E, Morency L-P (2016) Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259
90. Zhang Y, Qian X, Tan X, Han J, Tang Y (2016) Sketch-based image retrieval by salient contour reinforcement. IEEE Transactions on Multimedia (TOMM) 18(8):1604–1615
91. Zhang J, Shen F, Liu L, Zhu F, Yu M, Shao L, Tao Shen H, Van Gool L (2018) Generative domain-migration hashing for sketch-to-image retrieval. In: Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), pp 304–321
92. Zimmermann RS, Siems JN (2019) Faster training of mask R-CNN by focusing on instance boundaries. Computer Vision Image Understanding (CVIU) 188