Query-guided Attention in Vision Transformers for Localizing Objects Using a Single Sketch

Aditay Tripathi[1] Anand Mishra[2] Anirban Chakraborty[1]
[1]Indian Institute of Science [2] Indian Institute of Technology Jodhpur
{aditayt,anirban}@iisc.ac.in mishra@iitj.ac.in

Abstract

In this study, we explore sketch-based object localization on natural images. Given a crude hand-drawn object sketch, the task is to locate all instances of that object in the target image. This problem proves difficult due to the abstract nature of hand-drawn sketches, variations in the style and quality of sketches, and the large domain gap between the sketches and the natural images. Existing solutions address this using attention-based frameworks to merge query information into image features. Yet, these methods often integrate query features after independently learning image features, causing inadequate alignment and as a result incorrect localization. In contrast, we propose a novel sketch-guided vision transformer encoder that uses cross-attention after each block of the transformer-based image encoder to learn query-conditioned image features, leading to stronger alignment with the query sketch. Further, at the decoder's output, object and sketch features are refined better to align the representation of objects with the sketch query, thereby improving localization. The proposed model also generalizes to the object categories not seen during training, as the target image features learned by the proposed model are query-aware. Our framework can utilize multiple sketch queries via a trainable novel sketch fusion strategy. The model is evaluated on the images from the public benchmark, MS-COCO, using the sketch queries from QuickDraw! and Sketchy datasets. Compared with existing localization methods, the proposed approach gives a ***6.6****% and* ***8.0****% improvement in mAP for seen objects using sketch queries from QuickDraw! and Sketchy datasets, respectively, and a* ***12.2****% improvement in AP@50 for large objects that are 'unseen' during training. The code is available at https://vcl-iisc.github.io/locformer/.*

1. Introduction

Detecting objects in a natural image is an exciting area of research in computer vision. Notable progress in this domain over the past decade is mainly attributed to the continuous advancement of deep learning architectures [1, 18, 19, 24]. Despite these advancements, contemporary object detectors often face limitations in localizing instances of object categories not present during training. This issue poses challenges in practical scenarios, where a more versatile object localization technique is required to generalize effectively to unseen object categories—termed *open-world object localization*. One approach in this context is to localize objects in an image using an object image as a query [6]. However, this method may be hindered by the limited availability of object images due to copyright concerns, privacy restrictions, and data collection overhead, especially for uncommon or non-natural objects. Additionally, situations may arise where users lack access to query images of the desired object and may not even know the correct name. In these cases, they may choose to describe the object through a hand-drawn sketch. To gain insight into our motivation and objectives, please refer to Figure 1.

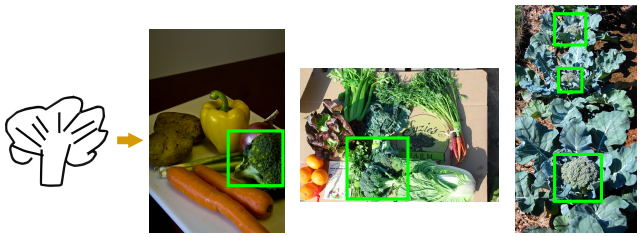

Figure 1. **Sketch-based object localization:** Consider a scenario where users wish to localize all the instances of the object *broccoli* on a set of natural images, and (i) images of *broccoli* are never seen during training, (ii) even at the inference time users do not have natural image of *broccoli* that can be used as a query, and (iii) the category name ("broccoli") is also unknown to the user. In such a situation, the user chooses to draw a sketch of *broccoli* by hand to localize all instances of it on the natural images. This work significantly improves the performance on this challenging task, namely sketch-guided object localization.

In addressing the aforesaid challenges, the problem of sketch-based object localization was introduced by Tripathi

et al. [25]. The primary goal is to localize all instances of an object on a natural image given a corresponding sketch query. However, this task presents significant challenges due to the abstract nature or 'crudeness' of the sketches, the diverse quality and style of hand-drawn sketches produced by non-expert users, and the domain gap between the query sketches and the target natural images. To tackle some of these obstacles, Tripathi et al. [25] proposed a cross-modal attention-based localization framework. Employing the Faster-RCNN architecture [19], they integrated a novel attention mechanism into the region proposal network (RPN), facilitating the generation of semantically relevant region proposals through the inclusion of a sketch-guided spatial attention score within the RPN. Subsequently, these region proposals undergo scoring to obtain the optimal object localization. Yet, this method heavily relies on the quality of the generated proposals, which might be rather limited, especially for occluded or underrepresented objects. Additionally, their method solely relies on spatial attention scores, possibly limiting localization performance due to the lack of explicit alignment between the query sketch and the target image features. Recently, Riba et al. [21] introduced Sketch-DETR, a novel extension of the popular object detection transformer (DETR) [1], achieving state-of-the-art performance in the sketch-based object localization task. Sketch-DETR employs an encoder-decoder transformer model, accepting both target image and sketch features obtained from their respective feature encoders as inputs. A self-attention mechanism is then applied to establish feature alignment between sketch and image features. However, a limitation of this methodology is rooted in its need for distinct image and sketch encoders to extract respective features before these are input into the self-attention mechanism, ultimately leading to sub-optimal alignment.

Addressing the limitations of current methods, we present a novel **sketch-guided vision transformer encoder**, which builds upon the vision and detection transformer (ViDT) [24]. Our sketch-guided encoder is designed to learn the representation of the target image conditioned on the query sketch, facilitating strong *feature alignment* between the image and the sketch. Specifically, our sketch-guided encoder takes the raw image as input, and after each block of the image encoder, a multi-headed cross-attention is applied to incorporate the query information into the image features. This process computes the attention score between image and sketch features, which is then used to fuse the sketch features into the image features. Consequently, the obtained target image features are more effectively aligned with the query sketch, enhancing query-guided localization performance.[1] Furthermore, after performing feature alignment within the sketch-guided encoder, we introduce *semantic alignment* at the decoder for further refinement. By utilizing multi-headed cross-attention between the object-level image features (represented by [DET] tokens) and sketch features, we semantically bring the representation of relevant objects closer to the sketch query representation, enabling precise and accurate localization. Thus, by proposing the novel feature and semantic alignment between the image and query sketch, we aim to overcome the limitations of existing methods and pave the way for more effective and accurate sketch-based object localization.

A key distinguishing aspect of our proposed framework is its robust performance under the challenging 'open-world' setting. Our framework demonstrates the ability to achieve highly accurate object localization, even for categories that are not part of the training data. It can be attributed to the effective alignment between the learned target image and the query sketch representations. Moreover, we introduce a trainable, novel sketch fusion strategy capable of combining complementary information from various sketches. The fusion process constructs a comprehensive object representation, significantly improving localization performance.

Contributions: In summary, we make the following contributions in this work: 1) To solve the sketch query based object localization task, we propose a novel sketch-guided vision transformer encoder that learns the representation of the target image conditioned on the query sketch, which leads to stronger alignment between the image and the sketch features and hence, much-improved localization accuracy. 2) Additionally, we propose a semantic alignment strategy at the output of the decoder that utilizes attention to bring the features of the relevant objects semantically closer to the sketch query, thereby further improving the localization performance. 3) We perform extensive evaluations on publically available benchmark datasets, where our proposed approach achieves a substantial gain of 8% over the best-reported results for sketch-based object localization on images from the MS-COCO dataset and query sketches from the Sketchy dataset. It, therefore, establishes a new state-of-the-art for this task.

2. Related Work

2.1. Object detection

Object detection is a well-studied yet open area of research in computer vision. Object detection approaches can broadly be grouped into (i) proposal-based and (ii) proposal-free methods. Although proposal-based methods [2,26] have several advantages, their performance is often limited by the quality of proposals they generate, which are often weak for occluded objects as well as 'unseen' object categories. Improving proposal generation for unseen

[1]Refer to suppl. materials for the comparison of our sketch-guided encoder with attention mechanisms proposed in [21, 25] using ViDT.

objects is an open area of research [31]. Our work falls under proposal-free methods. Among proposal-free methods, the modern transformer-based object detectors [1,24,32] are state-of-the-art. These methods often have encoder-decoder models and utilize a fixed set of [DET] tokens to learn to localize and classify the objects in the image. Methods such as ViDT [24] have progressed toward an encoder-free object detector, leading to fewer parameters and faster inference. While these object detection techniques are reasonably successful within the closed-world setting, there is a growing interest in addressing the challenges of open-world object localization [3, 6, 14]. In this space, [6] and [14] have proposed object localization in the one-shot setting and use the image of an object as a query. We address the problem of object localization for the scenario where the object category name is unknown, and the query image for the object of interest is unavailable; instead, a crude sketch representation of the object is available for the query to perform one-shot object localization. This recently introduced problem is referred to as *sketch-based object localization*.

2.2. Sketch-based Object Localization

Sketches have been applied to various computer vision tasks, e.g., sketch generation [8,16], 3D reconstruction [13], image and video retrieval [17, 28, 29], and 3D shapes retrieval [15, 27]. Recently, [25] introduced the problem of sketch-based object localization in natural images where, given a sketch query, the task is to localize the corresponding objects in the target images. They proposed a model based on Faster-RCNN and a cross-attention module to solve the problem. Yet, their model utilizes an inadequate attention mechanism that leads to subpar performance. Recently, [21] proposed a transformer-based approach, namely Sketch-DETR, where they concatenate the flattened sketch and image features before feeding them through the DETR encoder. Though more expressive than the cross-modal attention, they incorporate the sketch information at the encoder where target image features have already been learned. In this work, we propose *sketch-guided vision transformer encoder* that learns the representation of the target image conditioned on the query sketch by fusing the query information into the target image after each block of the transformer-based image encoder. Representation for the sketches has been learned using traditional CNNs [30], RNNs [4] and transformers [22] architecture. Our proposed framework requires the learned query representation regardless of the architecture used to learn the representation.

3. Methodology

3.1. Task Definition and Proposed Architecture

Consider a dataset $\mathcal{D} = \{I, S\}$ where I and S are sets of natural images and hand-drawn sketches, respectively. Let C be the set of all object categories in $\mathcal{D}$. Each image $I_i \in I$ contains bounding box annotations $B_i = \{(b_j, c_j)\}_{j=1}^{n_i}$ corresponding to all object instances (any of the C categories) present in it. Here, b_j is a rectangular box tightly surrounding the j^{th} object instance and $c_j \in C$ is the category of that object. Given a sketch $s \in S$ and an image $I_i \in I$, the problem of sketch-based object localization involves localizing all the instances of the object in the image that correspond to the sketch s. We address this problem in both *closed-world*, i.e., when object category of sketch query is 'seen' during training and *open-world*, i.e., when examples of the object category of sketch query are 'unseen' during training. An effective approach to tackle the task of sketch-based object localization involves extending the vision and detection transformer (ViDT) [24] to accommodate sketch queries. One plausible extension is to compare the representation of objects at the output of the decoder, denoted as [DET] tokens of ViDT, with the representation of the query sketch. Yet, this trivial extension might lead to suboptimal localization due to the independent learning of object and query features. In this study, we propose a novel *sketch-guided vision transformer encoder*, which addresses this limitation by learning the features of the target image conditioned on the query sketch. This conditional learning leads to better alignment between the target image representation and the query sketch, resulting in much-improved localization performance. Furthermore, we semantically refine the object and query features at the output of the decoder, aiming to bring the object features closer to the sketch query for better scoring. These refinements contribute to more precise localization results. The proposed model is end-to-end trainable and consists of three key components: (i) sketch-guided vision transformer encoder, (ii) object and sketch refinement, and (iii) scoring. The comprehensive architecture of our proposed model is illustrated in Figure 2.

3.1.1 Sketch-guided Vision Transformer Encoder

Traditional image encoders, such as ResNet [5], and modern transformer-based image encoders, such as Swin [12], consist of multiple layers of neural networks organized into blocks. Typically, an image is passed through all these blocks to learn its representation. Similarly, architectures like these can be employed to learn sketch embeddings independently. However, independently learned representations have poor alignment due to the substantial domain gap between the target natural image and the query sketch. To overcome this issue, we propose a novel approach that involves learning the representation of the target image conditioned on the query sketch. We achieve this by fusing the features of the sketch query with the image features at the output of each block of the transformer-based image encoder. In this work, we utilize the Swin-tiny transformer as

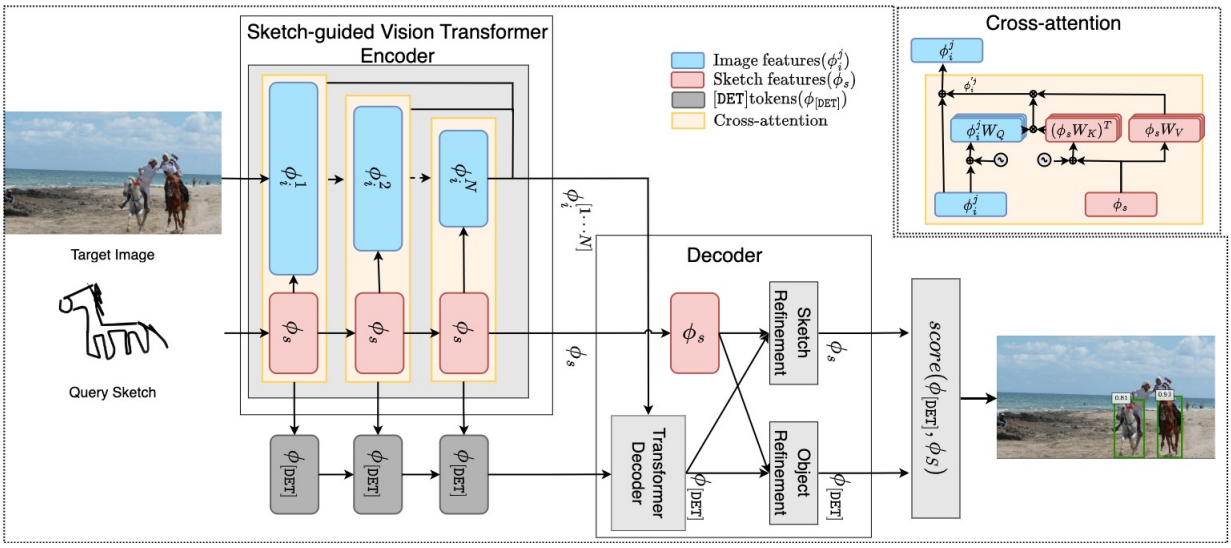

Figure 2. The proposed sketch-guided object localization model contains two primary components: (a) sketch-guided vision transformer encoder (Sec. 3.1.1) and (b) object and sketch refinement (Sec. 3.1.3). The sketch-guided transformer encoder takes the image at the input and generates sketch-conditioned features by fusing the sketch features into the target image after each block of the image encoder using **cross-attention**. After getting object-level features at the output of the **transformer decoder** ($\phi_{[DET]}$), the object features and the query sketch features are further refined to bring the features of the relevant object closer to the query sketch leading to better localization score.

our image encoder, which has demonstrated excellent performance in various computer vision tasks, and ResNet50 is used as the sketch encoder. Firstly, the features of the sketch query s are learned by passing it through a sketch encoder, and it is represented as $\phi_s \in \mathbb{R}^{d \times w \times h}$. Similarly, for the target image I_i, the representation at the output of the first block of the image encoder is given by $\phi_i^1 \in \mathbb{R}^{d \times w_I \times h_I}$.

Following each block in the image encoder, a pooling layer is employed to downsample the features, resulting in representations of the target image at different scales. We utilize multi-headed cross-attention to effectively fuse the sketch features into the image features at these different granularities. The image and sketch features are flattened before passing through the attention module. The image representation $\phi_i^1 \in \mathbb{R}^{w_I h_I \times d}$ is used as queries, and the sketch representation $\phi_s \in \mathbb{R}^{wh \times d}$ is used as key and value in the multi-headed cross-attention module. 2D sinusoidal position embeddings are added to the query and the key features to provide spatial features while calculating the attention weights. For brevity, we show the attention calculation for a single head, though we use multiple heads to learn the correspondence between the target image and the query sketch. The features of the target image are updated as:

$$\phi_i^{'1} = softmax\left(\frac{(\phi_i^1 \mathbf{W}_{Q_1})(\phi_s \mathbf{W}_{K_1})^T}{\sqrt{d'}}\right)\phi_s \mathbf{W}_{V_1}, \quad (1)$$

where d' is the dimension of the key vectors, $\mathbf{W}_{Q_1}, \mathbf{W}_{K_1} \in \mathbb{R}^{d \times d'}$ and $\mathbf{W}_{V_1} \in \mathbb{R}^{d \times d}$ is the projection matrices for the query, key, and value vectors respectively. These attended image features $\phi_i^{'1} \in \mathbb{R}^{w_I h_I \times d}$ are first transposed and further processed as follows $\phi_i^1 = \mathbf{W}_2\left(ReLU\left(\mathbf{W}_1 \phi_i^{'1}\right)\right) + \phi_i^1$. The image features ϕ_i^1 are reshaped to the original shape before feeding them through the next block, and the whole process is repeated for the remaining blocks in the image encoder. The similarity scores between the target image and the sketch features are first calculated in equation 1, and these scores are then used to fuse relevant sketch features into the target image features, leading to better alignment between the two modalities. The image features from each block of the sketch-guided vision transformer encoder are extracted and concatenated before being passed to the decoder. This concatenated representation, denoted as $\phi_i^{[1,\ldots,N]}$ in Figure 2, incorporates valuable information from all the blocks, enabling the decoder to benefit from query-aligned multi-scale image features and achieve more precise object localization results.

3.1.2 Object and Sketch Refinement

In our model, the decoder takes the image features at different scales to update the representation for the [DET] tokens. During training, these [DET] tokens are transformed into the representation of various objects at different locations within the image. Given that the decoder operates on sketch-conditioned image representations, the learned object features are better aligned with the query sketch. Importantly, the query fusion within the sketch-guided vision

transformer encoder occurs at a coarse-grained image feature level. We introduce a fine-grained fusion mechanism at the object/semantic level to further elevate performance. This fine-grained semantic fusion involves utilizing attention mechanisms to incorporate sketch features into the object features and vice versa. The proposed model brings the features of relevant objects closer to the query sketch, enabling more precise and contextually relevant localization.

Therefore, at the output of the decoder, the learned object features represented by the [DET] tokens and the sketch features are semantically refined to bring the representations of relevant objects closer to the query sketch. Given the flattened sketch representation $\phi_s \in \mathbb{R}^{wh \times d}$, the representations of [DET] tokens $\phi_{[\texttt{DET}]} \in \mathbb{R}^{100 \times d}$ are refined using multi-headed cross-attention as $\phi^{'}_{[\texttt{DET}]} = \phi_{[\texttt{DET}]} + \mathbf{W}_3 \left(ReLU \left(\mathbf{W}_4 \phi^{'T}_{[\texttt{DET}]} \right) \right)$, where $\phi^{'}_{[\texttt{DET}]}$ is obtained using similar attention computation defined in equation 1. 2D sinusoidal position encoding is added to the sketch representation before calculating the attention scores. Likewise, given the representation of [DET] tokens, the representation of the query sketch is refined by using a separate cross-attention module.

3.1.3 Scoring

After the semantic refinement, a scoring function Θ is learned to score each object instance on the image, given by [DET] tokens, against the query sketch. To do that, each of the [DET] tokens is assigned to each box in the ground truth using the Hungarian matching algorithm [7] as described in [1]. Each token ($[\texttt{DET}]_k$) is assigned a label y_k (1 or 0) based on whether it is assigned to a bounding box containing a foreground object, i.e., the object that corresponds to the sketch query. A global representation for the sketch is then obtained by taking the max-pool of the sketch feature maps, i.e., $\phi_S = \Psi(\phi_s)$, where $\phi_S \in \mathbb{R}^d$ and $\Psi : \mathbb{R}^{d \times w \times h} \rightarrow \mathbb{R}^d$.

Each of the [DET] token representations is concatenated with the global sketch representation before passing it through a neural network to generate the score for that token which is given by $score\left([\texttt{DET}]_k, s\right) = \sigma\left(\Theta\left([\phi_{[\texttt{DET}]_k}; \phi_S]\right)\right)$, where Θ is a neural network, σ is sigmoid function and $score : \mathbb{R}^{2d} \rightarrow [0, 1]$. The model is then trained to give high scores to the tokens that correspond to the objects in the query sketch by minimizing the following loss function:

$$L([\texttt{DET}], s) = \sum_k \Big\{ -y_k \ln\left(score([\texttt{DET}]_k, s)\right) - (1 - y_k)(1 - \ln\left(score([\texttt{DET}]_k, s)\right)) \Big\}. \quad (2)$$

Along with the classification loss defined in equation 2, regression loss and Generalized IoU [20] loss are also defined on the predicted bounding boxes with respect to the ground truth bounding box. During inference, all the [DET] tokens are scored with the query sketch, and the bounding boxes corresponding to high-scoring tokens are selected as the localized objects.

3.2. Multi-query localization

While our main focus is on achieving object localization through a single sketch query, we also explore the potential advantages of using multiple sketches for the same object to improve localization accuracy. Most hand-drawn sketches, such as those in the datasets used, are typically abstract and provide minimal information about object attributes and shapes. However, as highlighted in [25], employing multiple sketch queries can offer complementary information that enhances object localization in natural images. Prior techniques like *Feature fusion* and *Attention Fusion* proposed in [25] utilized max and mean pooling, respectively, for fusing information from multiple sketches, without incorporating any learnable element. In contrast, our work introduces a learnable query fusion approach, enabling our model to exploit information from multiple sketch queries more effectively and perform multi-query localization within our framework. While our method effectively handles single sketch queries (one-shot), we adapt Equation 1 to facilitate information fusion from multiple sketch queries. This empowers our model to harness the diverse information provided by multiple sketches, thereby enhancing localization performance.

$$\phi^{'1}_i[l] = S\left(\frac{\left(\phi^1_i \mathbf{W}_{Q_1}\right)\left(\phi_{s_l} \mathbf{W}_{K_1}\right)^T}{\sqrt{d'}} \right) \phi_{s_l} \mathbf{W}_{V_1}, \quad (3)$$

where, $\phi^{'1}_i[l]$ is the attention aggregated features for the l^{th} sketch query, and $S(.)$ is the softmax function. These features are aggregated for each query sketch, transposed, and then added to the image representations as follows:

$$\phi^1_i = \mathbf{W}_2 \left(ReLU \left(\frac{1}{L} \sum_{L=1}^{L} \mathbf{W}_1 \phi^{'1}_i[l] \right) \right) + \phi^1_i, \quad (4)$$

where L is the total number of query sketches.

At the decoder, we introduce an attention-based query fusion strategy to construct a unified sketch query representation from multiple sketches. Given n sketches and their corresponding feature map representations, we first calculate the average across the sketches to obtain the averaged feature map representation, denoted as $\phi_{s_\mu} \in \mathbb{R}^{d \times w \times h}$. Each query sketch representation is then flattened and concatenated, represented as $\phi_{s_{\{1,L\}}}$. Subsequently, we leverage attention mechanisms to incorporate complementary information present among the diverse sketches into the av-

Model	Query: Sketchy		Query: QuickDraw	
	mAP (%)	AP@50 (%)	mAP (%)	AP@50 (%)
Detection-based*				
FasterRCNN [19]	40.2	64.4	35.5	58.1
Retinanet [9]	42.2	66.1	37.9	60.1
DETR [1]	47.0	68.7	41.1	62.7
Localization-based				
Modified FasterRCNN [19]	-	-	18.2	31.5
CoAT [6]	-	-	27.9	48.6
CMA [25]	-	-	30.0	50.0
Sketch-DETR [21]	42.0	63.6	41.4	62.1
Ours	**50.0** (8.0 ↑)	**73.9** (10.3 ↑)	**48.0** (6.6 ↑)	**71.7** (9.6 ↑)

Table 1. Results in **closed-world, one-shot** setting. During inference, a single sketch from Sketchy and QuickDraw! respectively has been used as a query to localize 'seen' object categories on target images from *MS-COCO val2017* dataset, and mean average precision and AP@50 computed over all sketch queries have been reported. The numbers inside the parenthesis show gain with respect to the most competitive localization-based baseline. * Detection-based baselines assume the availability of a set of object categories.

eraged sketch representation. The attention-guided query fusion is defined as follows:

$$\phi_{s_\mu} = \phi_{s_\mu} + \mathbf{W}_5 \left(ReLU \left(\mathbf{W}_6 \phi_{s_\mu}^{'T} \right) \right), \tag{5}$$

$$\phi_{s_\mu}^{'} = S \left(\frac{\left(\phi_{s_\mu} \mathbf{W}_{Q_3} \right) \left(\phi_{s_{\{1,L\}}} \mathbf{W}_{K_3} \right)^T}{\sqrt{d'}} \right) \phi_{s_{\{1,L\}}} \mathbf{W}_{V_3}, \tag{6}$$

where $\mathbf{W}_{Q_3}, \mathbf{W}_{K_3} \in \mathbb{R}^{d \times d'}$ are the projection matrices for the query ϕ_{s_μ} and the key $\phi_{s_{\{1,L\}}}$ respectively, $\mathbf{W}_{V_3} \in \mathbb{R}^{d \times d}$ is the projection matrix for the value $\phi_{s_{\{1,L\}}}$, and $S(.)$ is the softmax function.

In our proposed method, the first step involves learning the correspondences between the average query feature (ϕ_{s_μ}) and all individual query features (ϕ_{s_i}, where $i = 1, \ldots, n$). These correspondences are then utilized to fuse complementary information present in the diverse sketches into the averaged sketch representation. This fused sketch representation is then used as the query in the refinement and the scoring stage. Moreover, $\mathbf{W}_i$ where $i = \{1, \cdots, 6\}$ are the learned projection matrices.

4. Experiments and Results

4.1. Datasets and Evaluation Setup

For our evaluations, we employed images from the MS-COCO dataset as target scenes while utilizing sketches from the QuickDraw! and Sketchy datasets as queries to access our model's performance. The Sketchy dataset [23] comprises 75,471 samples across 125 object categories. Each image in this dataset is paired with a crowd-drawn sketch, establishing a fine-grained image-sketch relationship. In contrast, the QuickDraw! dataset [4] contains an extensive collection of 50 million vector drawings across 345 object categories. We rasterized these sketches prior to inputting them into the sketch encoders. For target scenes, the MS-COCO dataset [11] was employed. This dataset shares 56 common categories with QuickDraw! and 27 common categories with Sketchy. To ensure evaluation consistency, we selected images with these shared categories from the COCO train2017 dataset and conducted evaluations on the COCO val2017 dataset.

Models	mAP	AP@50	AP^L
Modified FasterRCNN	3.3	7.4	6.2
CoAT [6]	5.9	12.4	10.6
CMA [25]	7.5	15.0	12.4
Ours	**12.2**	**18.3**	**24.6**

Table 2. Results in **open-world, one-shot** setting. Here, a single sketch query from QuickDraw! has been used to perform localization on 'unseen' categories of *COCO val2017* dataset. We observe that for large-sized objects, our approach outperforms state-of-the-art published results by 12.2% as measured by AP^L.

The performance evaluation takes place within two distinct setups: (i) *open-world one-shot* and (ii) *closed-world one-shot*. A single sketch is employed as a query to match the target image in both setups. Here, the term "one-shot" indicates the usage of a single sketch as the query. In the open-world scenario, 14 categories are excluded from the shared 56 categories between the QuickDraw! and MS-COCO datasets and labeled as 'unseen' categories. To strictly maintain the open-world setting, data from these 'unseen' categories is removed from Imagenet during pretraining as well. The sketch encoder is also pretrained

Model	Closed set				Open set			
	Sketchy		QuickDraw		Sketchy		QuickDraw	
	mAP	AP@50	mAP	AP@50	mAP	AP@50	mAP	AP@50
CMA	-	-	30.0	50.0	-	-	7.5	15.0
+ Query Fusion (5Q)	-	-	32.0	52.6	-	-	7.6	16.3
+ Feature Fusion (5Q)	-	-	32.0	53.1	-	-	8.0	17.1
Ours	50.0	73.9	48.0	71.7	15.1	23.9	12.2	18.3
+ Attention Fusion (5Q)	50.7	74.7	49.2	72.6	16.2	24.8	13.5	20.2

Table 3. Results in **multi-query** setting. Here, five sketch queries (5Q) are used to query images from *COCO val2017* dataset.

on QuickDraw! dataset with these 'unseen' categories excluded. Additionally, experiments under the 'multi-query' setting are conducted using a set of five sketch queries to query the target image.

4.2. Baselines

4.2.1 Detection-based baselines

In these baseline models, we adopt a methodology where the target image undergoes object detection to predict bounding boxes and their respective classes. Simultaneously, the query sketch is processed by a sketch classifier to predict its class. The predicted sketch category is then used to retrieve corresponding localizations from the object detectors' predictions. For comparison, we utilize three popular object detection models: FasterRCNN [19], Retinanet [10], and DETR [1]. For more details on training these methods, refer to [21]. It is important to note that these baseline methods presuppose the prior availability of object categories during evaluation. Consequently, they are only suitable for evaluation within a closed-world scenario, where object categories are known in advance.

4.2.2 Localization-based baselines

In these baselines, the sketch queries are directly compared with the representation of objects in the image to generate the localization. We used the following baselines in our experiments: **Modified FasterRCNN**: The region proposals are first generated using FasterRCNN. Then, the representation of region proposals is scored with the query sketch representation to obtain the localizations. For more details, refer to [25]. **Query-guided RPN**: In these baselines, the query information is incorporated in a region proposal network (RPN) to generate the region proposals relevant to the sketch query. To this end, we compared against two recent techniques, namely, CoATex [6] and cross-modal attention (CMA) [25]. **Sketch-DETR [21]:** uses a transformer-based object detector and concatenates the sketch query tokens with the image tokens at the input of the DETR encoder to incorporate the query information. This method has shown state-of-the-art results in sketch-based object localization.

4.3. Results and Discussion

4.3.1 Closed-world one-shot localization

Table 1, presents the results of sketch-based object localization for the closed-world, one-shot setting. Our model exhibits superior performance, outperforming the state-of-the-art methods substantially. The Modified Faster RCNN, which utilizes a query-independent region proposal network, performs poorly in comparison. Though CoAT and CMA, which employ query-dependent region proposal networks, show significant performance improvements, they still fall short compared to the detection-based methods and Sketch-DETR. The detection-based methods aim to bridge the domain gap between images and sketches by separately performing object detection and sketch recognition. They later map the predicted category of the sketch to the detected objects in the image. Yet, these methods are constrained by the performance of the object detectors and sketch classifiers, and most **importantly** require prior knowledge of the set of object categories. In contrast, Sketch-DETR uses a DETR-based localization framework, but the alignment between the sketch and image features is limited, resulting in relatively weaker performance. n contrast, our proposed sketch-guided vision transformer encoder facilitates query-aligned image features. Moreover, by incorporating semantic alignment at the decoder output, our model achieves state-of-the-art localization performance, with an **8%** improvement on the Sketchy dataset. This notable imporvement highlights our approach's efficacy, positioning it as a compelling solution for sketch-based object localization.

4.3.2 Open-world one-shot localization

The results for this challenging setting are presented for the MS-COCO val2017 dataset in Table 2. Sketch-DETR isn't included due to the absence of a public implementation. Our model shows superior performance, outperforming the current state-of-the-art method by **4.7**% in mAP and achieving an impressive **12.2**% improvement in $AP@50$ for large objects. The Modified FasterRCNN, using a standard region proposal network for 'unseen' objects, performs

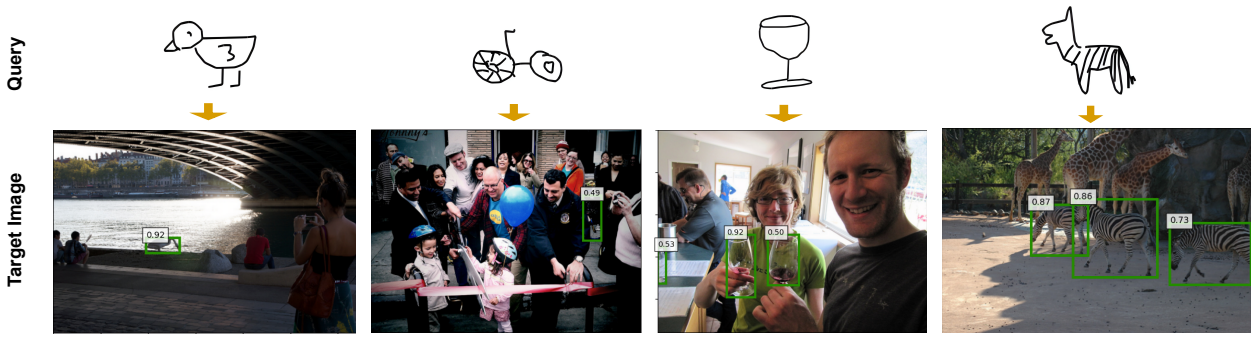

Figure 3. **Qualitative Results:** A selection of sketch queries and target images with queried objects localized using our proposed model are shown in the first and second rows, respectively. Our proposed model can localize occluded objects (such as *bicycle* in the second column) as well as multiple object instances (such as *glass* and *zebra* in the third and fourth columns) successfully. **[Best viewed in color]**

Model	mAP	AP@50
Modified-ViDT	39.4	56.6
+ Sketch-guided Vis. Trans.	46.9	68.7
+ Obj. and Sketch Refinement	**48.0**	**71.7**

Table 4. Effect of various components on the performance of the proposed model for the closed-world, one-shot setting.. The results are reported for COCO *val2017* images and queries from QuickDraw! dataset.

poorly. In contrast, CMA enhances performance with a query-dependent region proposal network. Our model outperforms both methods by adopting a more effective image and sketch alignment strategy, allowing generalized query-conditioned feature learning, even for unseen objects.

4.3.3 Multi-query localization

In this section, we investigate the impact of our attention-based query fusion strategy, detailed in Section 3.2, on multi-query object localization. The results, detailed in Table 3, utilize target images from the MS-COCO dataset and queries from both the QuickDraw! and Sketchy datasets. This experiment employs five diverse sketches as queries for the target image. Our proposed fusion strategy effectively integrates complementary information from different query sketches, leading to notable improvements in localization performance. Additionally, we present the multi-query localization results in the open-set scenario within Table 3. Notably, our query fusion strategy proves its effectiveness in elevating the performance of unseen categories.

4.3.4 Ablation

We examine the distinct contributions of each component to our model's performance, with corresponding results outlined in Table 4. In Vanilla-ViDT, sketch features are scored with [DET] tokens of ViDT for localization. The most substantial performance improvement is attributed to the query-aligned image features obtained at the output of the *sketch-guided vision transformer encoder*. This emphasizes the critical role of the feature alignment strategy, allowing the model to effectively leverage the query sketch's information for enhanced object localization accuracy. Additionally, the object refinement, implemented at the output of the decoder, further enhances the localization performance. This step introduces fine-grained semantic-level alignment, facilitating more precise localization results.

Qualitative Results: We perform a detailed qualitative analysis of our approach. A selection of results is shown in Figure 3. We observe that our method is successful in localizing occluded as well as multiple instances of objects. Further, we also did an error analysis of our approach and found that the major causes for failure are twofold: (i) ambiguity in sketch drawing, and (ii) highly overlapping objects. A detailed analysis is shown in the supplementary material.

5. Conclusion

In this work, we extensively studied the problem of sketch-guided object localization in natural images and proposed a novel transformer-based end-to-end trainable model. The proposed model uses the novel sketch-guided vision transformer encoder to learn sketch-conditioned image features. Further, object-level feature refinement at the decoder is performed to effectively align the natural image and the query sketch features. The effectiveness of the proposed model has been established by the state-of-the-art performance on publicly available benchmarks. Despite notable improvements in localization, sketch-based object localization still requires more research before deployment, we believe our work will inspire towards this purpose.

Acknowledgement This work is supported by a Young Scientist Research Award (Sanction no. 59/20/11/2020-BRNS) to Anirban Chakraborty from DAE-BRNS, India.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020.

[2] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[3] Akshita Gupta, Sanath Narayan, K. J. Joseph, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. OW-DETR: open-world detection transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9225–9234. IEEE, 2022.

[4] David Ha and Douglas Eck. A neural representation of sketch drawings. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[6] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-shot object detection with co-attention and co-excitation. In *NeurIPS*, 2019.

[7] Harold W. Kuhn. The hungarian method for the assignment problem. In Michael Jünger, Thomas M. Liebling, Denis Naddef, George L. Nemhauser, William R. Pulleyblank, Gerhard Reinelt, Giovanni Rinaldi, and Laurence A. Wolsey, editors, *50 Years of Integer Programming 1958-2008 - From the Early Years to the State-of-the-Art*, pages 29–47. Springer, 2010.

[8] Yi Li, Yi-Zhe Song, Timothy M. Hospedales, and Shaogang Gong. Free-hand sketch synthesis with deformable stroke models. *Int. J. Comput. Vis.*, 122(1):169–190, 2017.

[9] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society, 2017.

[10] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:318–327, 2020.

[11] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021.

[13] Zhaoliang Lun, Matheus Gadelha, Evangelos Kalogerakis, Subhransu Maji, and Rui Wang. 3d shape reconstruction from sketches via multi-view convolutional networks. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pages 67–77. IEEE Computer Society, 2017.

[14] Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection with vision transformers. *CoRR*, abs/2205.06230, 2022.

[15] Anran Qi, Yulia Gryaditskaya, Jifei Song, Yongxin Yang, Yonggang Qi, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. Toward fine-grained sketch-based 3d shape retrieval. *IEEE Transactions on Image Processing*, 30:8595–8606, 2021.

[16] Yonggang Qi, Guoyao Su, Qiang Wang, Jie Yang, Kaiyue Pang, and Yi-Zhe Song. Generative sketch healing. *Int. J. Comput. Vis.*, 130(8):2006–2021, 2022.

[17] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Deep shape matching. In *ECCV*, 2018.

[18] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[19] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[20] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 658–666. Computer Vision Foundation / IEEE, 2019.

[21] Pau Riba, Sounak Dey, Ali Furkan Biten, and Josep Lladós. Localizing infinity-shaped fishes: Sketch-guided object localization in the wild. *ArXiv*, abs/2109.11874, 2021.

[22] Leo Sampaio Ferraz Ribeiro, Tu Bui, John P. Collomosse, and Moacir Ponti. Sketchformer: Transformer-based representation for sketched structure. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 14141–14150. Computer Vision Foundation / IEEE, 2020.

[23] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016.

[24] Hwanjun Song, Deqing Sun, Sanghyuk Chun, Varun Jampani, Dongyoon Han, Byeongho Heo, Wonjae Kim, and Ming-Hsuan Yang. Vidt: An efficient and effective fully transformer-based object detector. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[25] Aditay Tripathi, Rajath R Dani, Anand Mishra, and Anirban Chakraborty. Sketch-guided object localization in natural images. In *ECCV*, 2020.

[26] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104:154–171, 2013.

[27] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1875–1883, 2015.

[28] Peng Xu, Kun Liu, Tao Xiang, Timothy M. Hospedales, Zhanyu Ma, Jun Guo, and Yi-Zhe Song. Fine-grained instance-level sketch-based video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 31:1995–2007, 2021.

[29] Qian Yu, Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Fine-grained instance-level sketch-based image retrieval. *Int. J. Comput. Vis.*, 129(2):484–500, 2021.

[30] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Sketch-a-net: A deep neural network that beats humans. *Int. J. Comput. Vision*, 122(3):411–425, may 2017.

[31] Ye Zheng, Ruoran Huang, Chuanqi Han, Xi Huang, and Li Cui. Background learnable cascade for zero-shot object detection. In Hiroshi Ishikawa, Cheng-Lin Liu, Tomás Pajdla, and Jianbo Shi, editors, *ACCV*, pages 107–123, 2020.

[32] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.