Appendix: 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
https://vcl-iisc.github.io/locformer/

A. Effect of dataset split

In the *open-world one-shot* setting, we conduct experiments by excluding a subset of randomly selected object classes from the training data. To assess the impact of dataset splits on localization performance, we performed experiments on four different data splits, and the results are reported in Table 1. Remarkably, our model achieves consistently high performance across all the data splits, with an average mean average precision (mAP) score of 12.1 ± 0.95. These results indicate the ability of the model to maintain a high level of performance across various dataset splits underscoring the generalization capabilities of our proposed approach, thereby enhancing its applicability in real-world settings with diverse and dynamically changing object categories.

B. Localization performance v.s. Object size

We evaluated the localization performance of our model for different object sizes, and the results are presented in Table 2. This analysis is performed in one-shot closed-set setting. Notably, our model exhibits high performance for medium and large-sized objects. Interestingly, when utilizing sketch queries from the Sketchy dataset, the model achieves significantly improved performance for small and medium-sized objects. This observation highlights the effectiveness of Sketchy dataset queries in handling smaller and medium-sized objects, further enhancing the model's localization capabilities across various object sizes.

C. Comparison of feature fusion methods

We compared the cross-modal attention (CMA) introduced in [2] and the self-attention (SA) used in [1] with the proposed sketch-guided vision transformer encoder. The ViDT architecture is utilized for this comparison, with results presented in Table 3. The CMA and SA are applied at the output of the ViDT feature extractor, and the sketch aligned image features are subsequently fed through the decoder. The learned [DET] are then scored with the sketch query to obtain the localization. In Modified-ViDT, the [DET] tokens are directly scored with the sketch query without any feature alignment step. The superior performance of our proposed sketch-guided vision-transformer encoder underscores its capability to establish stronger alignment between the target image and query sketch features, consequently enhancing localization accuracy. Conversely, CMA and SA methods did not exhibit comparable performance improvement. These observations underscore the efficacy of our sketch-guided vision transformer encoder, facilitating more effective alignment between the target image and query sketch features, ultimately leading to superior localization outcomes.

D. Robustness of query fusion technique

To evaluate the robustness of our proposed query fusion strategy concerning the number of query sketches, we conducted evaluations using our localization model trained on five sketch queries. We then evaluated the model's performance using two to eight sketch queries. The results, shown in Table 4, indicate that the model's performance remains consistent and does not vary significantly with the number of query sketches. Regardless of the number of query sketches used, the model maintains its effectiveness in localizing objects, further validating the reliability and adaptability of our approach.

E. Implementation Details

The proposed model is implemented using the PyTorch v1.9.1 library with CUDA 11.1 for GPU acceleration. Training is performed end-to-end on a single Nvidia-Quadro 8000 GPU (48GB VRAM) with a batch size of 7, the maximum batch size that our GPU can fit in its memory. We employ stochastic gradient descent (SGD) with a momentum of 0.9 as the optimization algorithm during training. The model is trained for 14 epochs, and the learning rate is set to $1e-5$ initially. After eight epochs, the

Split 1		Split 2		Split 3		Split 4		Average	
mAP	AP@50	mAP	AP@50	mAP	AP@50	mAP	AP@50	mAP	AP@50
12.2	18.3	13.1	19.4	10.8	16.9	12.2	18.1	12.1± 0.95	18.2 ± 1.02

Table 1. Results in **open-world** one-shot setting for different splits of the data. The sketches from the QuickDraw! dataset is used to query images from *COCO val2017* dataset.

Dataset	AP^S	AP^M	AP^L
Sketchy	17.9	47.1	69.3
QuickDraw	13.5	40.6	70.3

Table 2. The AP of the model for different sizes of the objects in the image.

Model	mAP	AP@50
Modified-ViDT	39.4	56.6
CMA-ViDT	42.3	63.5
SA-ViDT	43.0	66.8
Sketch-guided encoder	**46.9**	**68.7**

Table 3. Comparison of the proposed **sketch-guided vision transformer encoder** (*Sketch-guided encoder* in the table) with the attention mechanisms proposed in CMA and Sketch-DETR. The results are reported for queries from QuickDraw! dataset.

#Sketches	Sketchy		QuickDraw	
	mAP	AP@50	mAP	AP@50
2	50.5	74.6	47.8	70.6
3	50.6	74.6	49.0	72.5
4	50.8	74.8	49.1	72.5
5	50.7	74.7	49.2	72.6
6	50.8	74.7	49.2	72.7
7	50.8	74.7	49.3	72.8
8	50.8	74.8	49.3	72.8

Table 4. Performance of the model when the number of sketch queries changes during test time. The model trained on five sketches is used in this table.

learning rate is decayed by a factor of 0.1 to fine-tune the training process and improve convergence.

F. Computational details

On a Quadro RTX8000 GPU, our model takes an average of 55.2 ms per sample during inference, while the modified-ViDT without the Sketch-guided Vision transformer and Object and Sketch Refinement takes an average of 54.3 ms. However, the training time for both is around two days.

G. Additional qualitative results

We conducted a qualitative comparison of the localization performance between our model and the Cross-modal attention method proposed in [2]. The results are presented in Figure 1. Our model demonstrates the ability to accurately disambiguate between similar objects, as evident in the last column of Figure 1, where the model correctly localizes the object labeled as 'oven' in the image. In contrast, the Cross-modal attention method struggles to achieve the same level of accuracy for such cases. Furthermore, our proposed model excels in localizing multiple objects of the same type, even when they are in close proximity. For instance, in the 3rd and 4th columns of Figure 1, our model accurately localizes both the 'bear' and 'giraffe' objects, respectively. However, the Cross-modal attention method encounters difficulties in achieving precise localizations for such scenarios.

H. Failure case analysis

To underscore the complexities and challenges of the problem, we have included some of the failure cases in Figure 2. The model gives false positives in certain instances, particularly when the query sketch is highly ambiguous. For example, in the second row and second column of Figure 2, when presented with a sketch of a **dog**, the model erroneously localizes both the dog and the horse in the images, indicating the difficulty in disambiguating between similar objects. Additionally, the model encounters difficulties in scenarios where objects heavily overlap. For instance, in the second row and first column of Figure 2, when there are two zebras in the image, the model localizes only one of the objects, highlighting the challenges posed by densely overlapping objects. These failure cases underscore the intricacy of the sketch-based object localization task, where ambiguity in sketches and dense object arrangements can lead to inaccurate localizations.

References

[1] 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.

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

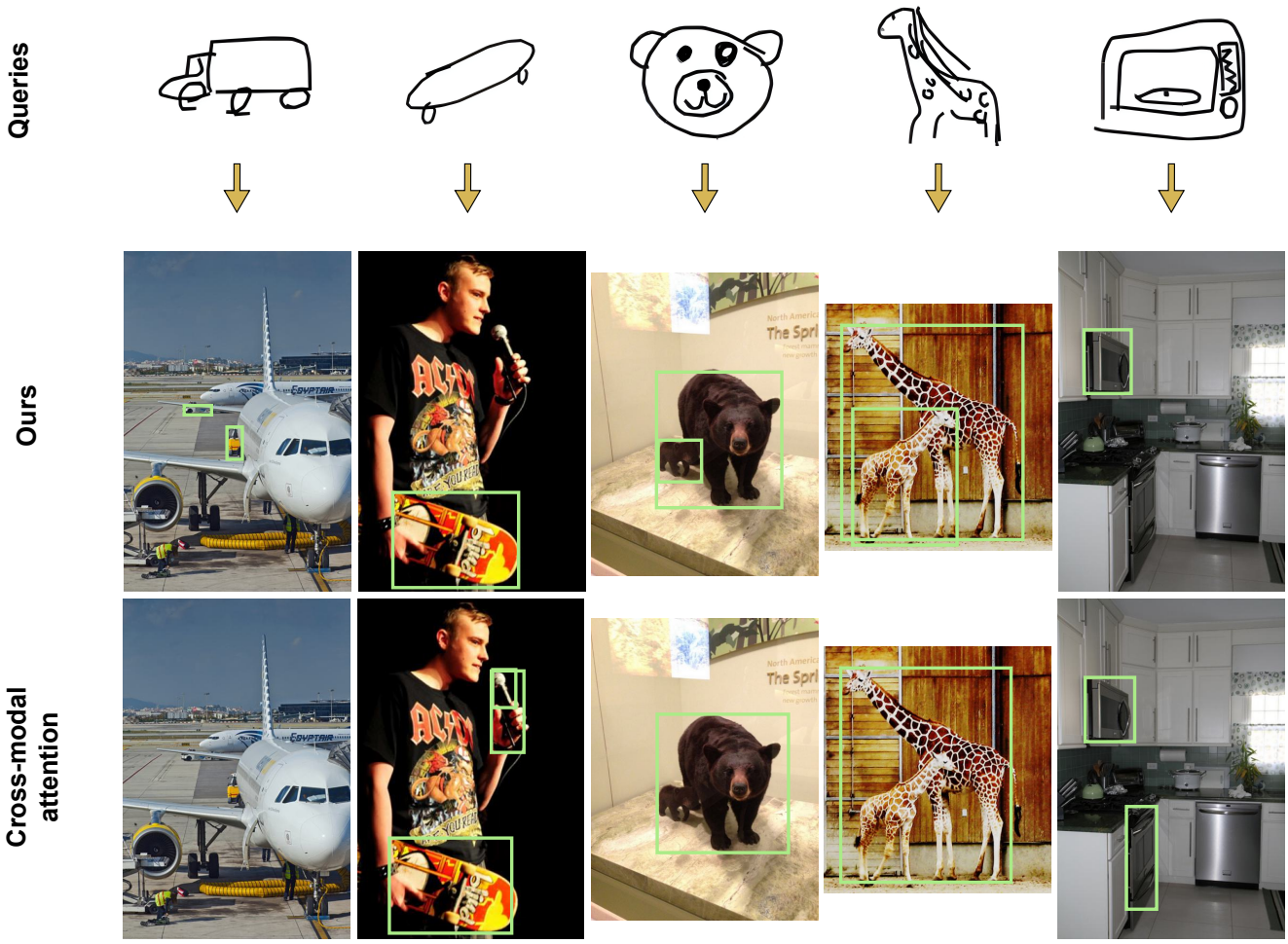

Figure 1. A selection of results comparing the previous work [2] i.e. cross-modal attention and this work. The first row shows the sketch queries. Green bounding boxes in the second and third rows show object localization using our work and previous work, respectively. [**Best viewed in color**].

Figure 2. We show some of the failure cases to highlight the challenges of the problem, e.g., in the second row and second column, the model is not able to disambiguate the correct object because the query sketch itself is very vague.[**Best viewed in color**].

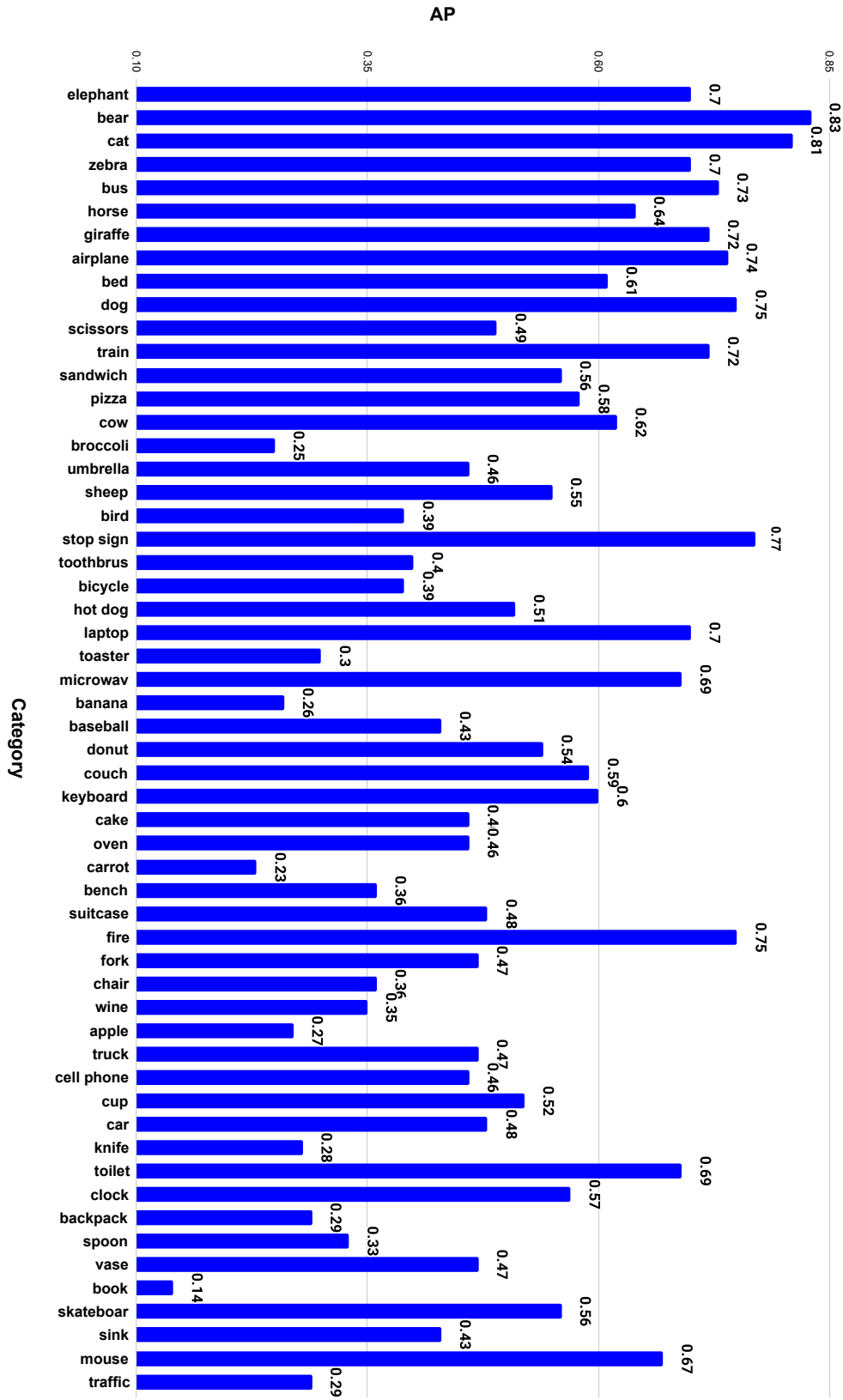

Figure 3. **Class-wise** %**AP** results on images from MS-COCO dataset and sketches from QuickDraw! Dataset.