VISUAL QUERIES 2D LOCALIZATION TASK

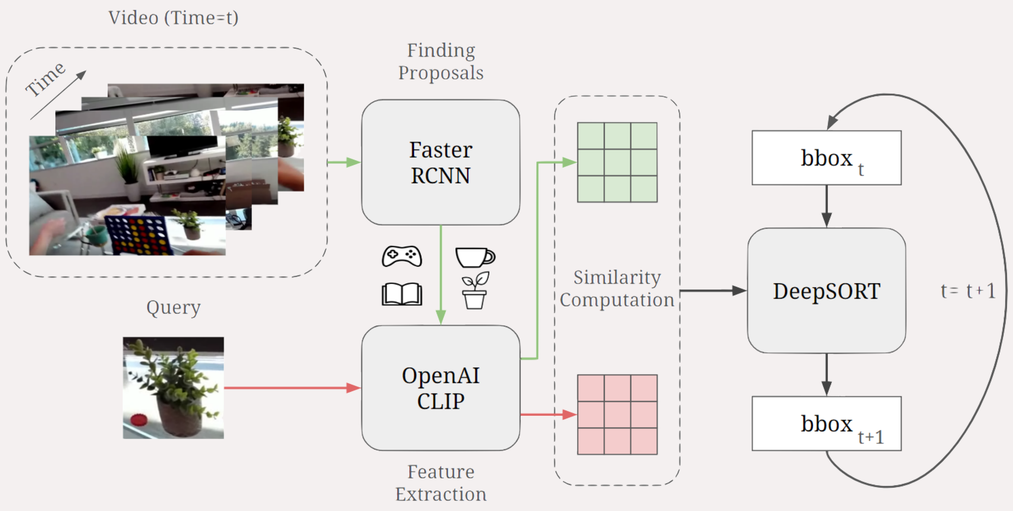
Group 11 Boss'sDog: 土木所二 程懷恩 **R11521701,** 土木所二 何宏發 **R11521616**
森林所二 許致銓 **R10625016,** 電信所二 黃湛元 **R11942180**

Introduction

In this project, we implemented 2 methods to achieve visual query localization. One is the traditional two-stage solution including object detection and frame-wise object tracking, we used Faster R-CNN and DeepSORT. The other is reproducing the state-of-the-art VQLoC, a single-stage transformer-based model, utilizing 3-D positional encoding to increase the performance of spatial-temporal transformer.

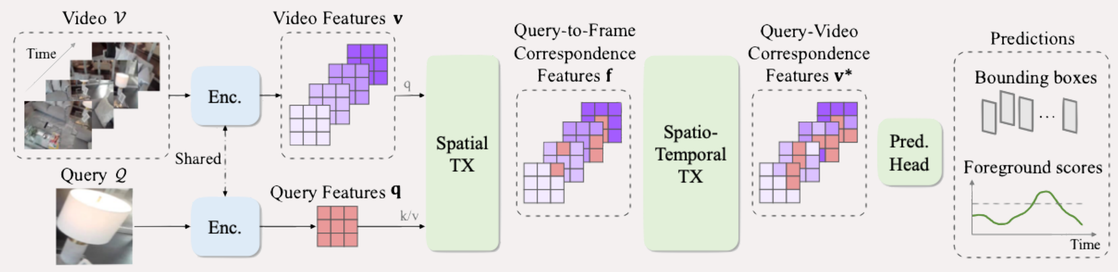
FASTER R-CNN + DEEPSORT

We leveraged the high-precision object detection capabilities of Faster R-CNN, coupled with the stable tracking of DeepSORT. Faster R-CNN efficiently identifies objects bounding boxes as region proposals within video feeds, while DeepSORT, utilizing Kalman filters and a Convolutional Neural Network, persistently tracks these objects.



VQLOC

VQLoC focused on training an end-to-end visual query localization framework. Reinforced by the ViT model, VQLoC can better generalize the long-range query-to-frame correspondence. Furthermore, VQLoC also generalizes the frame-to-frame relationship with a temporal window for ViT. Besides, avoiding the similarity calculation of the proposal and the query, VQLoC inferences 36 FPS as state-of-the-art.



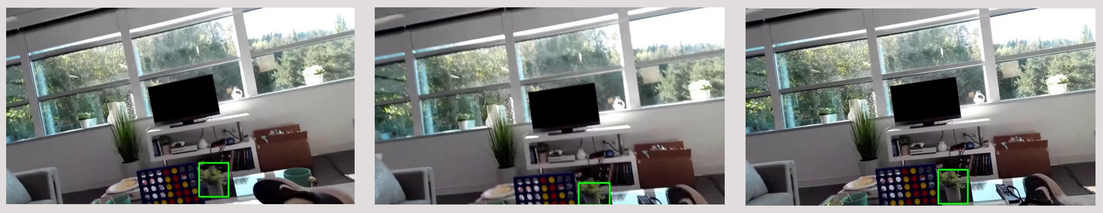
Model Architecture of VQLoC[3]

Based on VQLoC pre-trained model, we attempted to adjust the small modules in the algorithm:

1. **The kernel size of median filter**
2. **The peak window threshold**
3. **Normalization of query and frames**

Data Visualization

Ground Truth

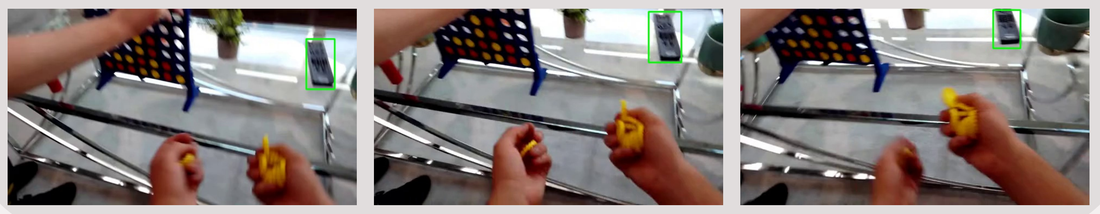


Prediction

Data Visualization

Ground Truth

Prediction



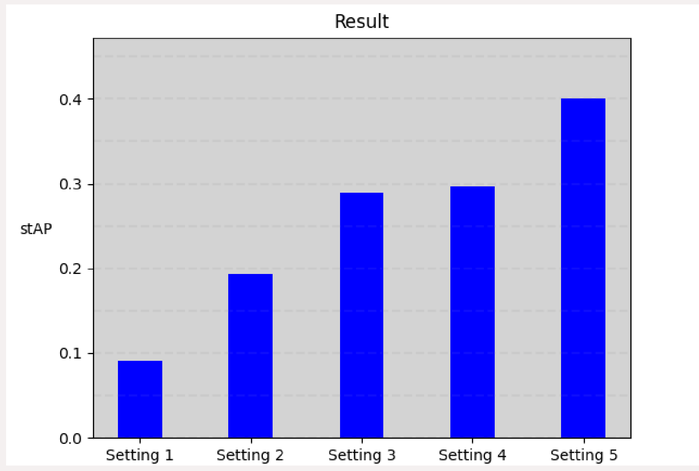
Results & Ablation Study

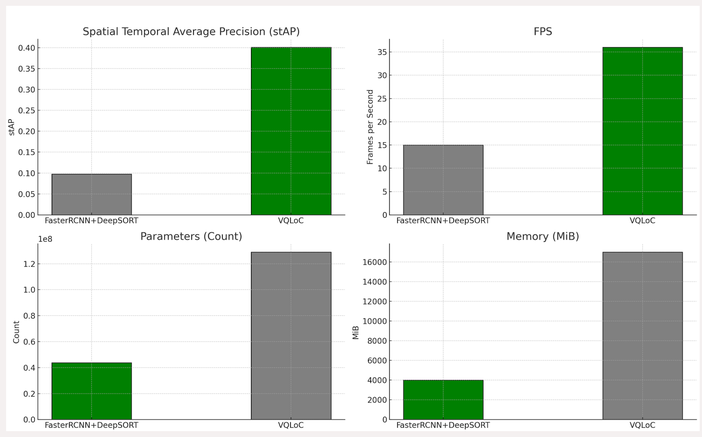
First, we implemented Faster R-CNN and DeepSORT. We start with Faster R-CNN to find proposals, then filter them by comparing the feature similarity with the query. Finally, the recognized results are fed into DeepSORT for predicting and tracking the movement of objects using a Kalman filter.

Second, we tested three different settings in VQLoC pretrained model with the former method.

1. **Faster R-CNN + DeepSORT (Setting 1)**
2. **Regular query and frames (Setting 2)**
3. **Normalized query and frames (Setting 3)**
4. **Setting 3 + smaller median filter kernel (Setting 4)**
5. **Setting 4 + prediction score (Setting 5)**

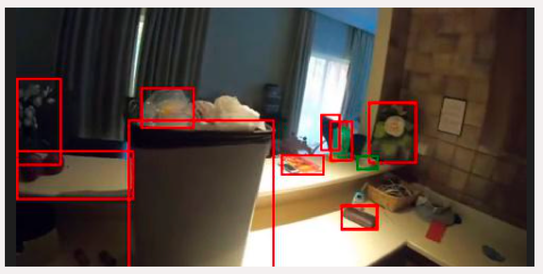
We found that the normalized frame-to-frame data and smaller kernel size can provide us with a more generalized result, leading to the stAP 0.2965, while the regular query and frames can only achieve stAP 0.1938. In our opinion, the normalized frames can be more consistent by normalization, and a smaller kernel size can contribute to the more fine-grained features. We added the query prediction scores to increase the performance based on the calculation of stAP.





Conclusion

Based on the two mainstream solutions for the visual query localization problem, we found that the state-of-the-art VQLoC undoubtedly performs better, somehow even more precisely than our labels. Nevertheless, through the implementation of a two-stage framework, we detailedly understood the step-by-step method to localize the target objects. Besides, we realized that the most critical part lies in regional proposals. Without the proposals including our target object, it is difficult to process and predict afterward even if we include a reliable tracking method.



REFERENCES

[1] Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple Online and Realtime Tracking. In: ICIP, pp. 3464–3468. IEEE (2016)
[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NeurIPS, 2015
[3] Jiang, H., Ramakrishnan, S. K., & Grauman, K. (2023). Single-Stage Visual Query Localization in Egocentric Videos. arXiv preprint arXiv:2306.09324.
[4] Xu, M., Li, Y., Fu, C.-Y., Ghanem, B., Xiang, T., & Pérez-Rúa, J.-M. P. (2023). Where is my Wallet? Modeling Object Proposal Sets for Egocentric Visual Query Localization. Meta AI, KAUST, Saudi Arabia.
[5] Wojke, N., Bewley, A., & Paulus, D. (2017). Simple Online and Realtime Tracking with a Deep Association Metric. University of Koblenz-Landau, Queensland University of Technology.