VQ-PL: Visual Query Localization in Egocentric Videos with Pseudo Label

周奕節（R10943131）林孟平（R12943010）謝宗翰（R12921A10）張祐綸（R12943122）

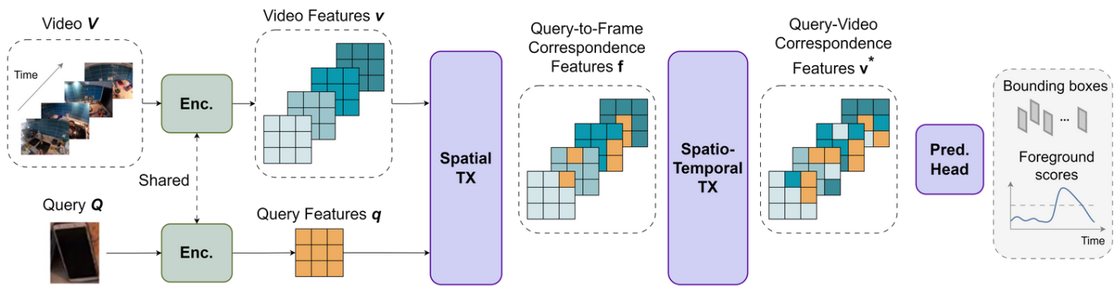
NTU DLCV-FALL-2023 LOLIDYULUNLOVEYUKI

INTRODUCTION

Visual Query Localization on long-form egocentric videos requires spatio-temporal search and localization of visually specified objects and is vital to build episodic memory systems. We reproduced the idea of VQLoC and used some data augmentation and training techniques to further improve the performance of the original model. The result shows that our approach surpasses baseline by a wide margin. The concept can leave more exploration for future work and be generalized to other tasks.

METHODOLOGY

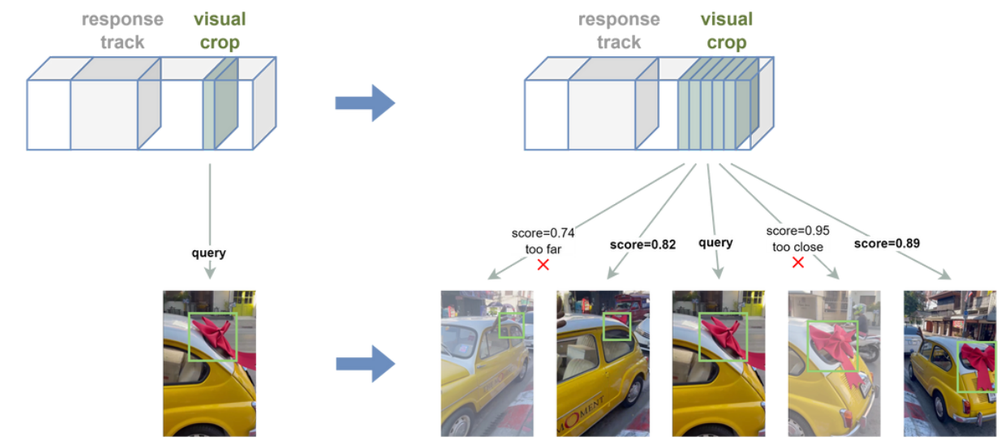
I. MODEL



II. DATA AUGMENTATION

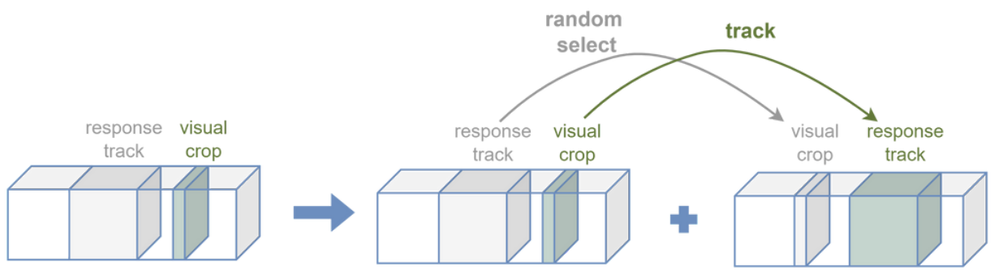
a. Unlabeled Frames Sampling as Query (UFS-Q)

Collect additional queries from frames near the visual crop with confidence score near 0.8.



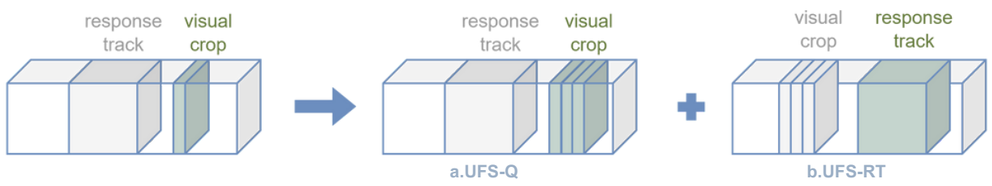
b. UFS as Response Track (UFS-RT)

Randomly select a response track as a new visual crop and track the visual crop to get a new response track.

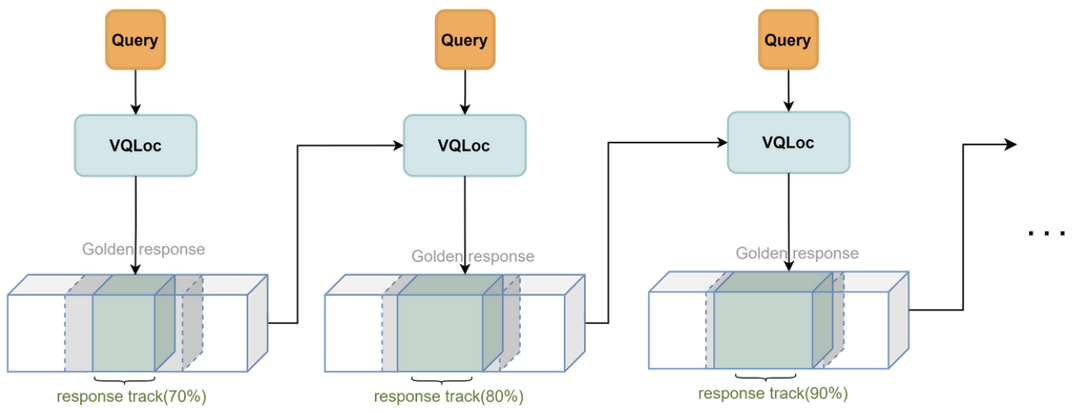


c. Mixed UFS-Q and UFS-RT

Combine the above two methods to get more diverse visual crops and response tracks.



III. RECURSIVE PSEUDO LABEL (RPL)



RESULT

I. EXPERIMENT (ABLATION STUDY)

Comparison between different UFS-Q settings:

Augmentation	Val IoU	Val Probability	Val stAP	Test stAP
None	0.504	**0.875**	0.295	0.328
UFS-Q (3)	0.504	0.817	0.290	0.341
UFS-Q (5)	**0.514**	0.806	**0.317**	**0.346**

Comparison between different UFS-RT settings:

Augmentation	Val IoU	Val Probability	Val stAP	Test stAP
None	0.504	**0.875**	0.295	0.328
UFS-RT (10)	0.519	0.867	0.299	**0.343**
UFS-RT (20)	**0.552**	0.856	**0.322**	0.330

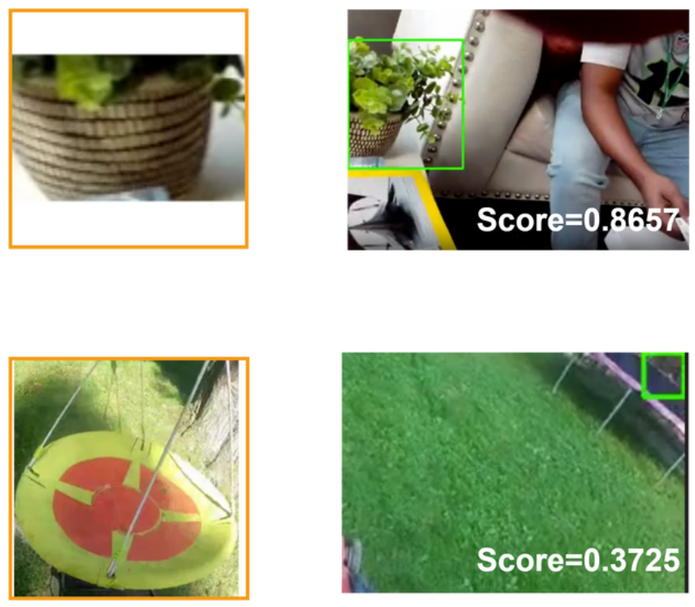
Comparison between UFS-Q and UFS-RT:

Augmentation	Val IoU	Val Probability	Val stAP	Test stAP
None	0.504	**0.875**	0.295	0.328
UFS-Q	0.504	0.817	0.290	**0.346**
UFS-RT	**0.519**	0.867	**0.299**	0.343

Comparison between all settings (Modified Inference):

Augmentation	Val IoU	Val Probability	Val stAP	Test stAP
None	0.504	**0.875**	0.332	0.379
UFS-Q	0.514	0.806	0.330	0.388
UFS-RT	**0.520**	0.867	0.331	0.383
UFS-RPL	0.519	0.814	**0.340**	**0.395**

II. OUTPUT VISUALIZATION



CONCLUSION

Our experiment result shows that our approach can indeed improve the performance of the model (0.2897 → 0.3945). UFS and UFS-Q can produce various visual crops and additional response tracks. The model can thus see different aspects of the query object and generate a more accurate bounding box. However, as the method of Pseudo Label, UFS-Q and UFS-RT still face the problem of mislabeling. In future work, we plan to refine these Pseudo Label methods, complete the recursive steps of RPL and do more research on the original model architecture.

BIBLIOGRAPHY

H. Jiang, et al., "Single-Stage Visual Query Localization in Egocentric Videos." in 2023 Conference on Neural Information Processing System (NeurIPS).

M. Xu, et al., "Where is my Wallet? Modeling Object Proposal Sets for Egocentric Visual Query Localization." in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023 pp. 2593-2603.

M. Xu, et al., "Negative Frames Matter in Egocentric Visual Query 2D Localization." arXiv preprint arXiv:2208.01949, 2022.