Visual Queries 2D Localization (VQ2D)

GROUP BASELINER

陳韋傑	r12922051
魏子翔	r12922147
劉憶欣	r11525085
趙祖翎	r11525131

ABSTRACT

In this project, we have selected the DINOv2 Visual Large-Scale Linguistic Model as our backbone. It allows us to efficiently capture image features and implement spatial and temporal localization. The entire model is structured in the form of a transformer, enabling improved information transfer and interaction. This structure is particularly effective in handling the spatial and temporal localization of picture features during the training process. Consequently, our model can flexibly address the interaction of visual and linguistic information.

METHODOLOGY

- **Modify the model structure:**

Model 1 is the original structure presented in this paper, comprising the process of downsampling and upsampling the image after the encoder. Downsampling leads to the loss of small object features. As a point of reflection, the following transformations are applied to the model:

  - Model 2: Combine with a 16*16*256 feature.
  - Model 3: Combine with a 16*16*256 temporal transfer feature.

- **DIoU loss:**

DIoU introduces a distance metric that takes into account the distance between the predicted box and the actual target box.

$$DIoU = IoU - \frac{\rho^2(b, b^{gt})}{c^2}$$

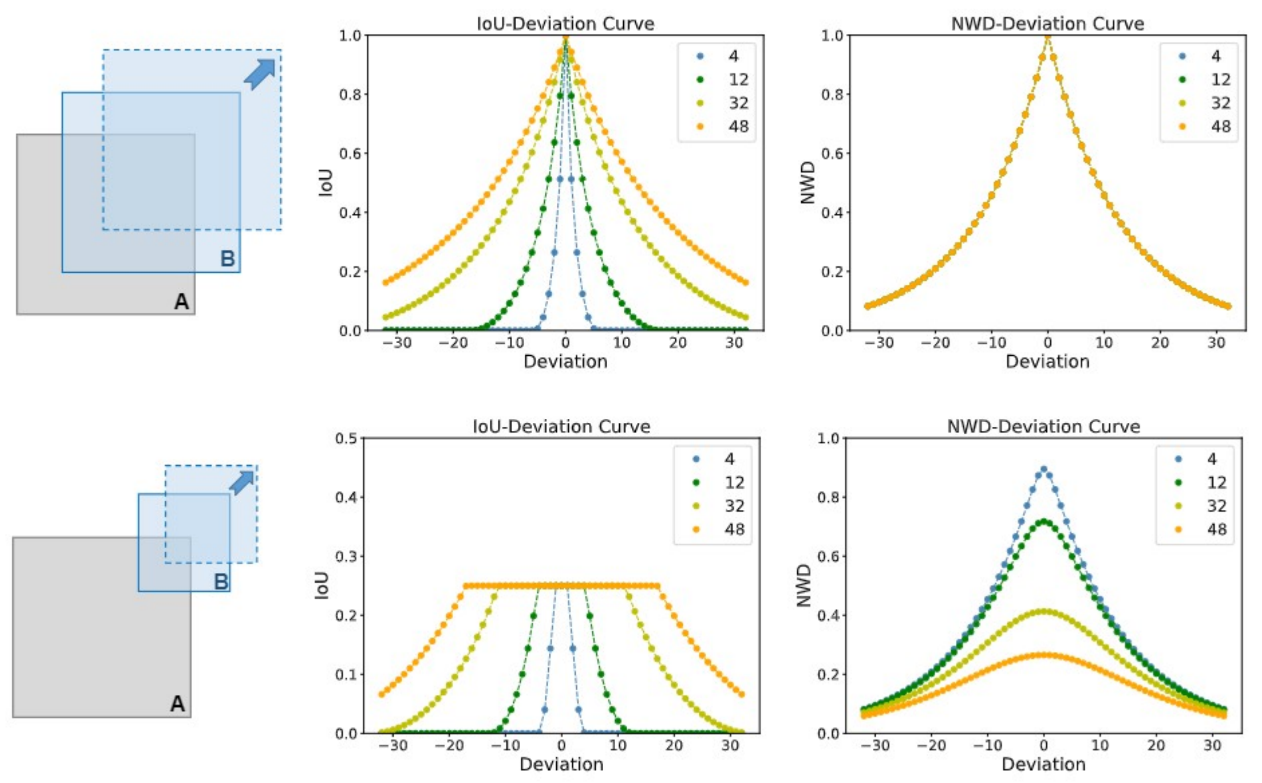
- **NWD loss:**

NWD employs Wasserstein distance to measure the similarity of Bounding Boxes as a replacement for the standard IoU. For two Bounding Boxes, the Wasserstein distance is computed :

$$W_2^2(N_a, N_b) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2}\right]^T, \left[cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2}\right] \right) \right\|_2^2$$

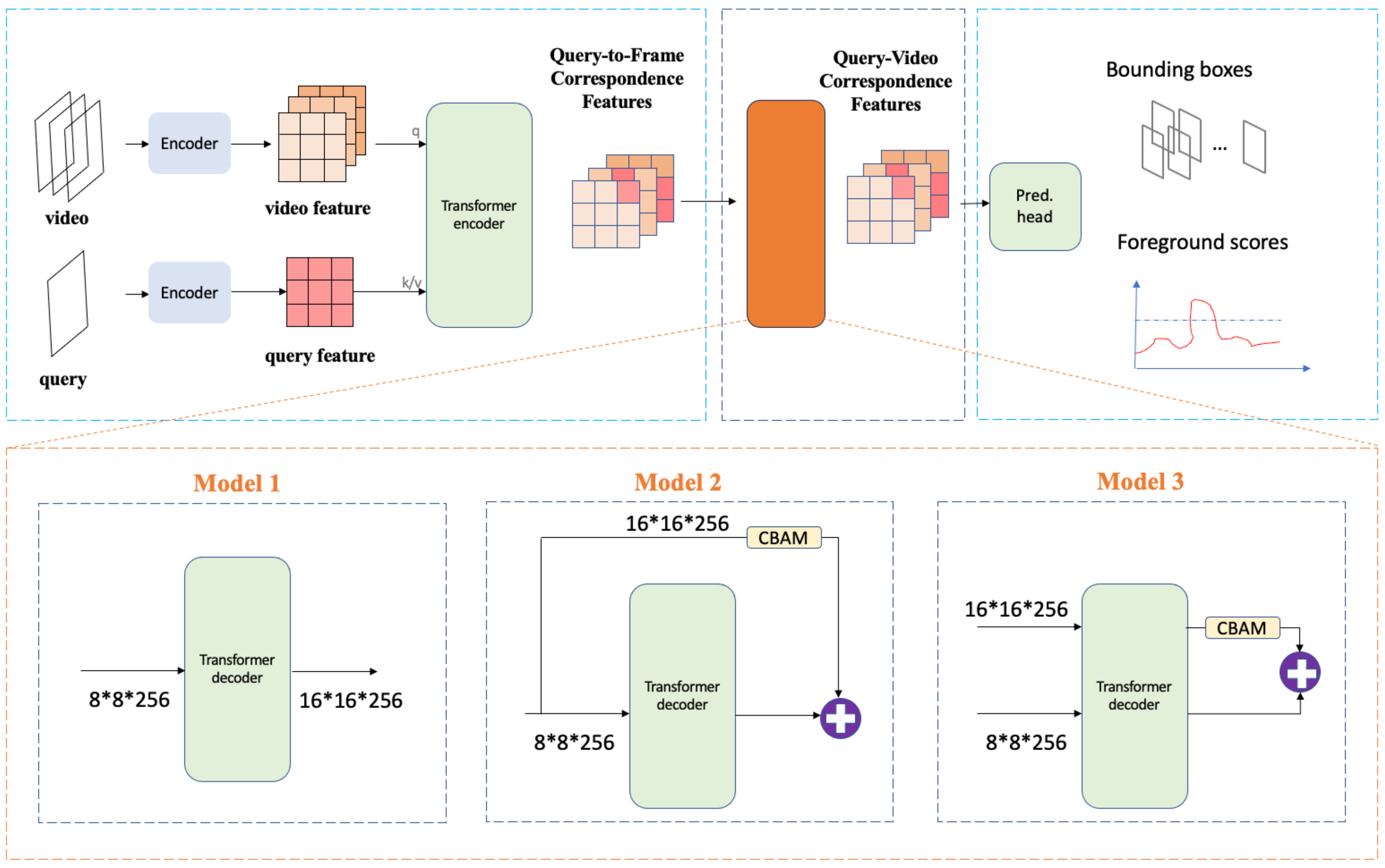
Exponentiating and normalizing it, we obtain the Normalized Wasserstein Distance (NWD) as a metric for similarity:

$$NWD(N_a, N_b) = \exp\left(-\frac{\sqrt{W_2^2(N_a, N_b)}}{C}\right)$$

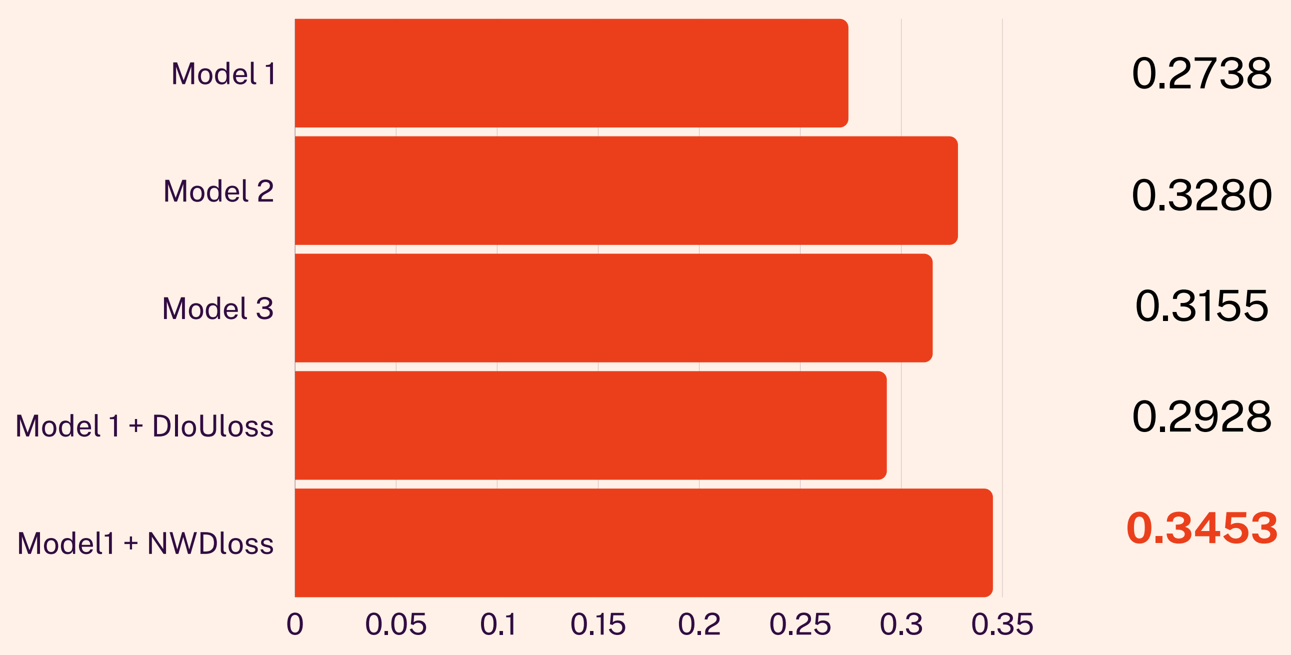


INTRODUCTION

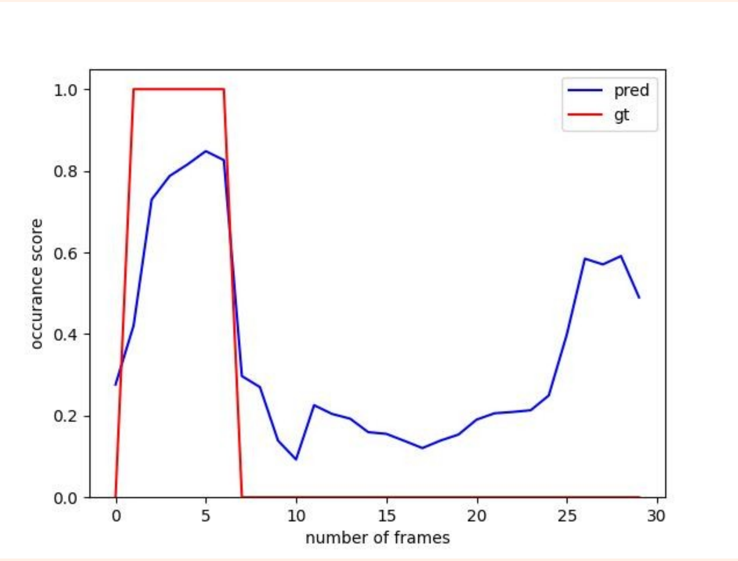
Based on the VQLOC model structure, a notable limitation is the relatively inaccurate detection of small objects. Consequently, in the construction of our own model, our primary emphasis lies in enhancing the detection of small objects through the fusion of multi-scale spatial fidelity and the integration of NWD loss.



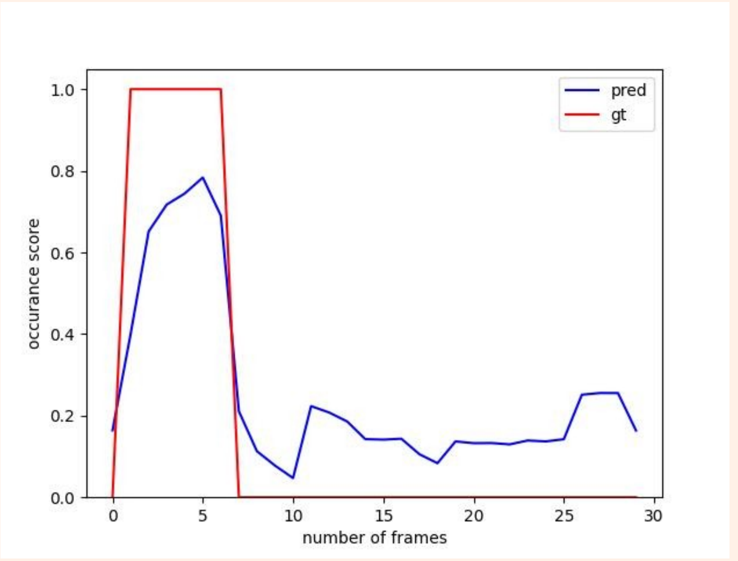
ANALYSIS



Model 1 with pretrained



Model 1 with NWD loss



From the analysis of the previous figures, it becomes evident that the pretrained model exhibits two peaks in its predictions, whereas the NWD model shows only one peak. Upon scrutinizing the video and query image, it is affirmed that the second peak in the pretrained model prediction is a false positive (the queried object does not appear in the video). Therefore, we conclude that incorporating NWD in the loss function proves beneficial in mitigating issues related to small object detection.

CONCLUSION AND FUTURE WORK

We conducted an extensive exploration of diverse model architectures tailored to tackle the intricacies of the Visual Query Localization (VQL) problem in egocentric videos. In our pursuit to elevate the detection accuracy of smaller targets, we strategically employed feature fusion on multi-scale feature maps and transitioned the loss function to the sophisticated Normalized Wasserstein Distance (NWD). As a testament to the efficacy of these enhancements, the stAP performance on Codalab yielded a notable achievement, reaching a commendable score of 0.3453.

REFERENCE

1. HanwenJiang, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Single-stage visual query localization in egocentric videos .arXiv preprint arXiv:2306.09324,2023.
2. Wang, J.; Xu, C.; Yang, W.; Yu, L. A normalized Gaussian Wasserstein distance for tiny object detection. arXiv 2021, arXiv:2110.13389.
3. Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-IoU loss: faster and better learning for bounding box regression, arXiv:1911.08287.
4. Woo, S.; Park, J.; Lee, J.; Kweon, I. S. CBAM: Convolutional block attention module. In: Computer Vision — ECCV 2018. Lecture Notes in Computer Science, Vol. 11211. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 3–19, 2018.