Transformer-based 3D Single Object Tracker

Akshay Kumar Sureddy Roshini Pulishetty

April 6, 2024

Project description. We plan to implement 3D single object tracking on LIDAR point clouds. The main objective of this project is to localize the given object in the dynamic search space at every time step. This method takes 3D LIDAR point clouds as input and estimates the location of the bounding box center (x, y, z) and its up-scale orientation θ.

Single-object tracking has been instrumental in many scenarios, like autonomous driving, robotics, and surveillance systems. 2D single object tracking algorithms relied heavily on RGB data of images, which degraded their performance in dark or illumination-changing environments. Unlike in 2D, LIDAR point clouds are widely used in 3D. They are not only robust to illumination and weather conditions but can also capture accurate geometric information, thus providing reliability in a range of driving conditions. However, these point clouds are sparse and disordered hence mere CNN processing is insufficient. They are often incomplete and irregular, making this problem challenging.

Problem Statement. Given the initial state of the target in a dynamic 3D scene, the tracker locates the target frame-by-frame in the video sequence. That is, by accessing the initial bounding box $\mathcal{B}_0$ of the target and the past and the current frames $\{P_i\}_{i=1}^{t}$, the model finds the current bounding box $\mathcal{B}_t$. The bounding box $\mathcal{B}_t \in R^7$ contains coordinates of the center (x, y, z), size (l, w, h) and orientation with up-axis (θ). Assuming the bounding-box size is a non-varying component and the object is ground-aligned, requiring only 1 rotation component, we predict (x, y, z, θ) in each frame for 4 degrees of freedom.

Overview of architecture. Our architecture is based on a Siamese-based paradigm, that encompasses 1. A Siamese Network that encodes feature seeds for template and search area. 2. Transformer to augment features. 3. A regression-based 3D Target Proposal and Verification network that outputs the bounding boxes. 4. (Optional) RNN-based motion forecasting component.

Related work. Recently, the Siamese network-based paradigm garnered attention in 2D object tracking. Developing similar networks for point clouds, P2B [1] proposed a pipeline, where PointNet++ backbone extracts seeds from template and search areas. Then, it embeds template seeds into the search area by permutation-invariant feature augmentation. Consequently, a joint 3D target proposal and verification network regresses the potential target centers via Hough Voting to output the bounding box. Later, PTT [2] emphasized that a transformer module plays a non-trivial role in the voting and proposal generation phases.

Further, a few studies leveraged transformers for feature representation and voting in their architecture. One such architecture is GLT-T [3], which proposes a Global-Local Transformer as a voting scheme. Moreover, PTTR [4] and OSP2B [5] supported incorporating self and cross-attention for the fusion of template and target-specific seeds.

In our work, we plan on re-implementing some components from P2B and OSP2B, to introduce a novel architecture that harnesses Siamese shared backbone for feature extraction and the power of transformers in feature fusion.

Datasets. We will train and evaluate our model on the KITTI dataset. We compare our results against the P2B baseline.

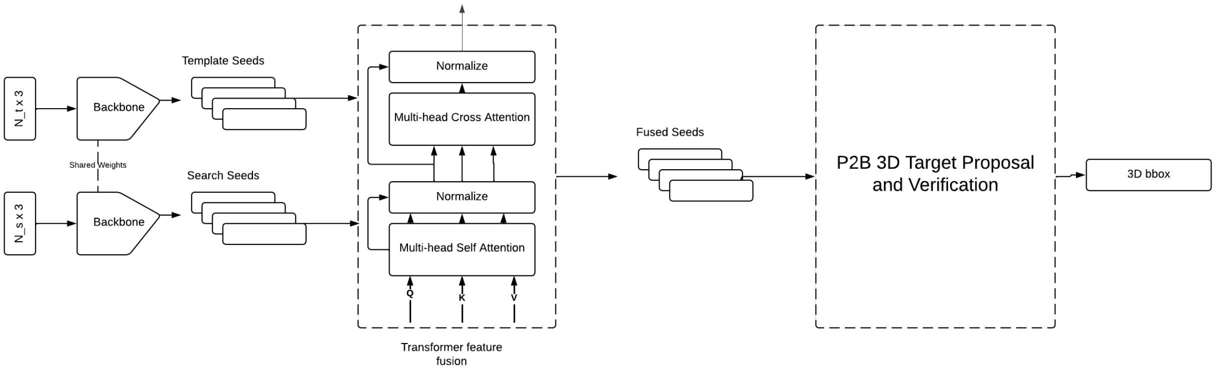
N_t x 3
Backbone
Shared Weights
N_s x 3
Backbone
Template Seeds
Search Seeds
Normalize
Multi-head Cross Attention
Normalize
Multi-head Self Attention
Q
K
V
Transformer feature fusion
Fused Seeds
P2B 3D Target Proposal and Verification
3D bbox

Figure 1: Overview of Transformer-based Tracker Architecture

Evaluation Metric. We employ success and precision metrics using One Pass Evaluation. Success is the intersection over the Union (IoU) between the predicted and ground-truth bounding box, that is, the overlap of the predicted box with the ground-truth box. Precision is the distance of the center of the predicted bounding box from the ground-truth box, with a threshold up to 2 meters.

Split of the work. We plan on splitting the tasks equally and working collectively by discussing updates weekly. The below tasks and their assignee are tentative and may change depending on the requirement-

- Data Pre-processing (Akshay)
- Implementing data loaders (Akshay)
- Implementing feature extraction module (Roshini)
- Implementing feature fusion module (Akshay)
- Implementing 3D target proposal and verification (Roshini)
- Training and hyper-parameter tuning. (Roshini)
- Ablation study to analyze the importance of each component, robustness to varying numbers of proposals, task alignment scores, and effectiveness of seed-wise targetness score. (both)
- Developing and evaluating against baselines. (both)
- Qualitative analysis of the running speed of video (in fps), advantageous cases, and error-prone scenarios. (both)
- (Optional) Implement architecture for predicting the target in the future frame using RNNs or Transformers. (both)

References

[1] Haozhe Qi, Chen Feng, Zhiguo Cao, Feng Zhao, and Yang Xiao. P2b: Point-to-box network for 3d object tracking in point clouds, 2020.

[2] Jiayao Shan, Sifan Zhou, Zheng Fang, and Yubo Cui. Ptt: Point-track-transformer module for 3d single object tracking in point clouds. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1310–1316, 2021.

[3] Jiahao Nie, Zhiwei He, Yuxiang Yang, Mingyu Gao, and Jing Zhang. Glt-t: Global-local transformer voting for 3d single object tracking in point clouds, 11 2022.

[4] Changqing Zhou, Zhipeng Luo, Yueru Luo, Tianrui Liu, and Liang Pan. Pttr: Relational 3d point cloud object tracking with transformer, 2022.

[5] Jiahao Nie, Zhiwei He, Yuxiang Yang, Zhengyi Bao, Mingyu Gao, and Jing Zhang. Osp2b: One-stage point-to-box network for 3d siamese tracking, 2023.