3D Prompt Learning for RGB-D Tracking Supplementary Materials

Bocen Li[1], Yunzhi Zhuge[1], Shan Jiang[2], Lijun Wang[1]*, Yifan Wang[1], and Huchuan Lu[1]

[1] Dalian University of Technology, Dalian, China
[2] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China

1 Detail Design of Geometry Prompt Block

In Tab. A, we analyze the impact of different design choices within Geometry Prompt (GP) block. We also provide results without the score branch to only investigate the impact of our method on the transformer backbone.

The base model MixFormer operates at different scales, namely 1/4, 1/8, and 1/16, with varying numbers of transformer blocks, specifically 1, 4, and 16. It can be seen that using $[1, 2, 8]$ geometry prompt blocks at the corresponding scales outperforms the other configuration. Our experiments with different configurations of GP block in the base model MixFormer reveal that applying geometry prompt block before every transformer block is not the optimal strategy. This observation highlights the importance of carefully designing and integrating 3D information into the tracking framework to strike the proper balance between 2D and 3D features.

Additionally, considering that the original point clouds may contain noise, refining the geometric point features f_{geo}^{3D} using confidence scores also leads to improved performance.

Table A. The impact of detailed design choices in the proposed GP block. GP Blocks refer to the number of these blocks used at the respective scale. Confidence indicates whether the point cloud confidence is utilized to refine the geometry point features.

GP Blocks	Confidence	F(↑)	Pre(↑)	Rec(↑)
$[1, 2, 4]$	✔	0.597	0.585	0.610
$[1, 2, 8]$	✔	0.615	0.602	0.628
$[1, 4, 16]$	✔	0.600	0.588	0.612
$[1, 2, 8]$	✘	0.599	0.587	0.611

2 Inference Speed

We compare ours (OSTrack_3DPT and MixFormer_3DPT) with the most recent state-of-the-art RGB-D methods on DepthTrack. All comparisons are conducted under the same setting with an NVIDIA 3090 GPU. As shown in the

* Corresponding author.

Tab. B, our inference speeds are near or beyond real-time and comparable with other methods. Considering the tracking accuracy, we believe that our methods have achieved a better trade-off between performance and speed.

Table B. FPS of different RGB-D tracking algorithms.

Method	ARKitTrack	ViPT	OSTrack_3DPT	MixFormer_3DPT
FPS	27.48	39.38	28.7	21.47

3 Point Cloud Visualizations

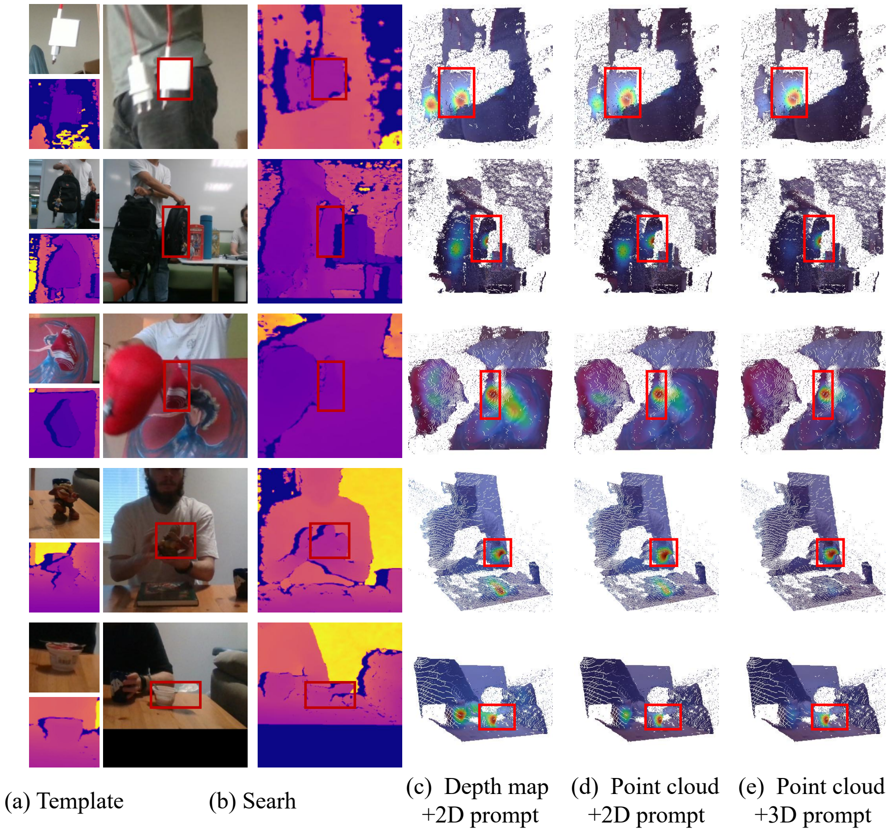

Fig. A. Visualizations in 3D space of different methods.

To further investigate the importance of incorporating 3D information, we extend Fig. 5(a) from the paper by visualizing the score maps on point clouds, as shown in Fig. A. In this figure, targets are highlighted by red bounding boxes. In the 3D space, it becomes clearer that our method, denoted as (e), can accurately localize targets.