JUNE 18-22, 2023
CVPR VANCOUVER, CANADA

# Referring Multi-Object Tracking

Dongming Wu[1*], Wencheng Han[2*], Tiancai Wang[3], Xingping Dong[4], Xiangyu Zhang[3,5], Jianbing Shen[2+]

[1]Beijing Institute of Technology, [2]SKL-IOTSC, University of Macau, [3]MEGVII Technology,

[4]Wuhan University, [5]Beijing Academy of Artificial Intelligence

Poster: WED-PM-218

Project Page: https://referringmot.github.io/

# RMOT overview


The parking cars


The red cars



The cars which are turning


The cars in the counter direction of ours


The cars in left



The black cars which are moving


The pedestrian


The persons in the right



The cars which are slower than ours

- We propose a new referring understanding task (RMOT) that uses a language expression to detect and track multiple objects.
- We formulate the first RMOT benchmark based on KITTI, called Refer-KITTI.
- We provide a baseline model based on Transformer, named TransRMOT.

# Motivation

- In previous referring understanding benchmarks, each expression tends to ground only one target, lacking simulation on the multi-object scenarios.
- Besides, the given expression only describes part of frames for the video referring task, making the correspondence inaccurate.

To solve these problems:

- We propose a novel video understanding task guided by the language description, named **referring multi-object tracking (RMOT)**.
- Given a language expression as a reference, it targets to ground all semantically matched objects in a video, including **multiple referent targets and temporal status variances**.

(a) Query: the cars in the right

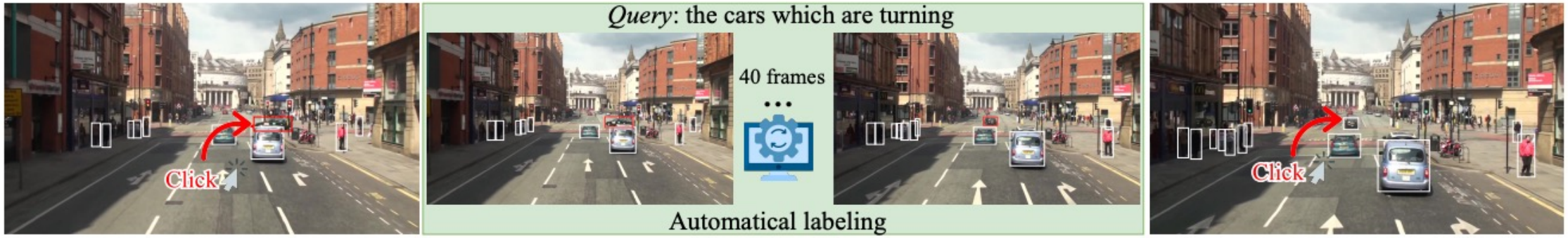
(b) Query: the cars which are turning

Figure 1. **Representative examples from RMOT**. The expression query can refer to multiple objects of interest (a), and captures the short-term status with accurate labels (b).

# Contributions

- We propose a new task for referring multi-objects, called referring multi-object tracking (RMOT). It tackles limitations in the existing referring understanding tasks and provides multi-object and temporally status-variant circumstances.
- We formulate a new benchmark, Refer-KITTI, to help the community to explore this new field in depth. As far as we know, it is the first dataset specializing in an arbitrary number of object predictions
- We propose an end-to-end framework built upon Transformer, termed as TransRMOT. With powerful cross-modal learning, it provides impressive RMOT performance on Refer-KITTI compared to hand-crafted RMOT methods.

# Benchmark




- Dataset Annotation with Low Human Cost:
  - We make full use of the instance-level box annotations of KITTI and design an efficient labeling tool.
  - For example, the turning action is labeled with only two clicks on bounding boxes at the starting and ending frames. The intermediate frames are automatically labeled with the help of unique identities.

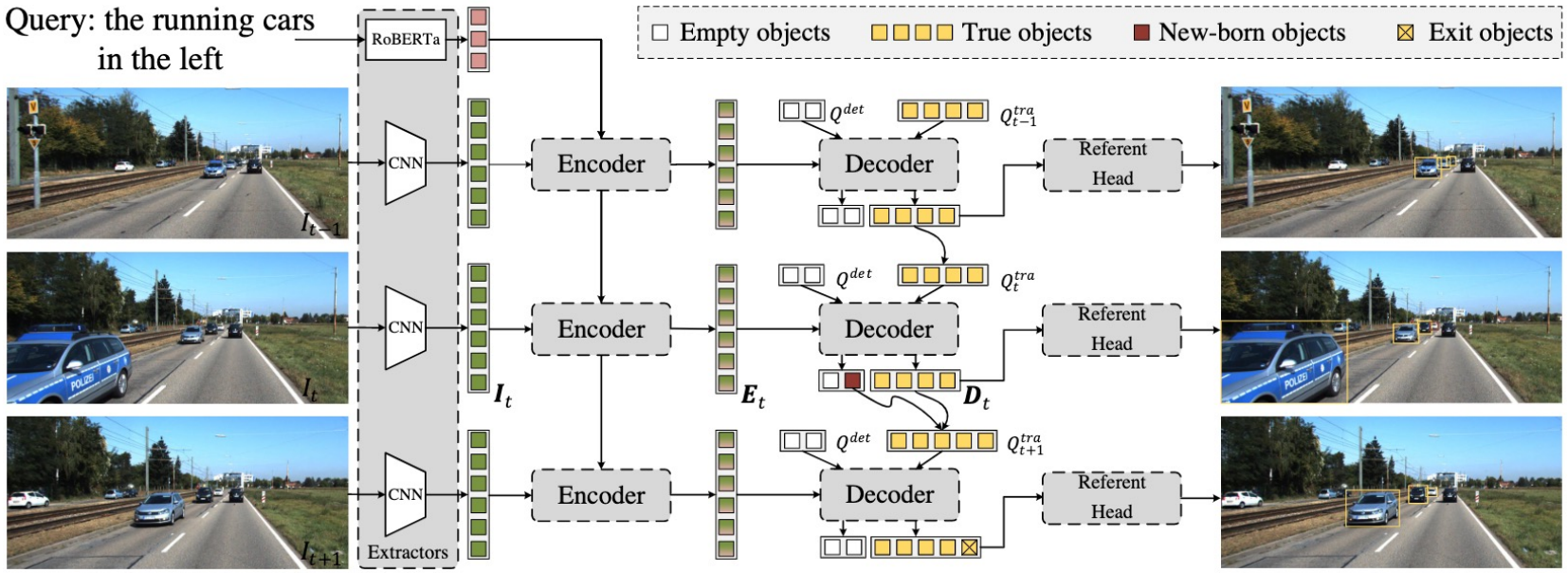
# Benchmark

- Dataset Features and Statistics:
  - **High Flexibility with Referent Objects.** Different from previous datasets that contain just one referent object for each language expression, RMOT is designed to involve an arbitrary number of predicted objects in videos.
  - **High Temporal Dynamic.** The per-expression temporal ratio covering the entire video indicates many referent objects enter or exit from visible scenes.
- Evaluation Metrics:
  - Higher Order Tracking Accuracy (HOTA) is used as standard metrics to evaluate the new benchmark. Its core idea is calculating the similarity between the predicted and ground-truth tracklet.

| Dataset | Video | Images | Instances per-expression | Temporal ratio per-expression |
|---|---|---|---|---|
| RefCOCO [55] | - | 26,711 | 1 | 1 |
| RefCOCO+ [55] | - | 19,992 | 1 | 1 |
| RefCOCOg [55] | - | 26,711 | 1 | 1 |
| Talk2Car [5] | ✓ | 9,217 | 1 | - |
| VID-Sentence [4] | ✓ | 59,238 | 1 | 1 |
| Refer-DAVIS$_{17}$ [17] | ✓ | 4,219 | 1 | 1 |
| Refer-YV [38] | ✓ | 93,869 | 1 | 1 |
| Refer-KITTI | ✓ | 6,650 | 10.7 | 0.49 |

Table 1. **Comparison of Refer-KITTI with existing datasets.** Refer-YV is short for Refer-Youtube-VOS. The temporal ratio represents the average ratio of referent frames covering the entire video sequence. '-' means unavailable.

# Method




- We propose an end-to-end differentiable framework for RMOT, named TransRMOT
- It builds upon the DETR framework, enhanced by powerful cross-modal reasoning and cross-frame conjunction
  - We modify the encoder as a cross-modal encoder for fusing the features of two modalities.
  - We employ track query to associate same objects between adjacent frames

# Experiments

| Method | HOTA (△HOTA) | DetA | AssA | DetRe | DetPr | AssRe | AssPr | LocA |
|---|---|---|---|---|---|---|---|---|
| FairMOT [59] | 22.78(± 0.87) | 14.43 | 39.11 | 16.44 | 45.48 | 43.05 | 71.65 | 74.77 |
| DeepSORT [45] | 25.59(± 0.79) | 19.76 | 34.31 | 26.38 | 36.93 | 39.55 | 61.05 | 71.34 |
| ByteTrack [58] | 24.95(± 0.84) | 15.50 | 43.11 | 18.25 | 43.48 | 48.64 | 70.72 | 73.90 |
| CStrack [21] | 27.91(± 0.73) | 20.65 | 39.10 | 33.76 | 32.61 | 43.12 | 71.82 | 79.51 |
| TransTrack [40] | 32.77(± 0.68) | 23.31 | 45.71 | 32.33 | 42.23 | 49.99 | 78.74 | 79.48 |
| TrackFormer [31] | 33.26(± 0.65) | 25.44 | 45.87 | 35.21 | 42.19 | 50.26 | 78.92 | 79.63 |
| TransRMOT (Ours) | **35.54**(± 0.71) | **28.25** | **46.25** | **39.22** | **45.94** | **50.69** | **80.67** | **79.79** |

- Our model TransRMOT achieves promising detection and tracking performance quantitively compared to other counterparts.

# Experiments

Query: the black cars in the left

Query: the cars which are faster than ours

Query: the pedestrian in the right

- TransRMOT successfully predicts referent objects according to the given expression.

# Thanks!