Mono3DVLT: Monocular-Video-Based 3D Visual Language Tracking

Hongkai Wei[1*], Yang Yang[1*], Shijie Sun[1†], Mingtao Feng[2†], Xiangyu Song[1†], Qi Lei[1], Hongli Hu[1], Rong Wang[1], Huansheng Song[1], Naveed Akhtar[3], Ajmal Saeed Mian[4]
[1] Chang'an University, China [2] Xidian University, China [3] The University of Melbourne, Australia
[4] University of Western Australia, Australia

Abstract

*Visual-Language Tracking (VLT) is emerging as a promising paradigm to bridge the human-machine performance gap. For single objects, VLT broadens the problem scope to text-driven video comprehension. Yet, this direction is still confined to 2D spatial extents, currently lacking the ability to deal with 3D tracking in the confines of monocular video. Unfortunately, advances in 3D tracking mainly rely on expensive sensor inputs, e.g., point clouds, depth measurements, radar. Absence of language counterpart for the outputs of these mildly democratized sensors in the literature also hinders VLT expansion to 3D tracking. Addressing that, we make the first attempt towards extending VLT to 3D tracking based on monocular video. We present a comprehensive framework, introducing (i) the **Mono**cular-Video-based **3D V**isual **L**anguage **T**racking (**Mono3DVLT**) task, (ii) a large-scale dataset for the task, called Mono3DVLT-V2X, and (iii) a customized neural model for the task. Our dataset is carefully curated, leveraging a Large Langauge Model (LLM) followed by human verification, composing natural language descriptions for 79,158 video sequences aiming at single object tracking, providing 2D and 3D bounding box annotations. Our neural model, termed Mono3DVLT-MT, is the first targeted approach for the Mono3DVLT task. Comprising the pipeline of multi-modal feature extractor, visual-language encoder, tracking decoder and a tracking head, our model sets a strong baseline for the task on Mono3DVLT-V2X. Experimental results show that our method significantly outperforms existing techniques on the Mono3DVLT-V2X dataset. Our dataset and code are available in https://github.com/hongkai-wei/Mono3DVLT.*

1. Introduction

Object tracking [36, 37, 47], the capability to visually follow objects over time, is a fundamental human skill that has inspired extensive research in vision. Recently, efforts have intensified to develop tracking models that emulate human tracking behaviors [11, 12]. Nevertheless, despite progress, Single Object Tracking (SOT) [10, 42] remains limited by the constraints of unimodal approaches, which impede human-like perception and restrict its applications in complex tasks. Visual Language Tracking (VLT) [19] has recently emerged as a promising framework, adding language as an auxiliary modality to enhance perceptual depth. However, existing VLT methods are limited to 2D spatial extents and lack the capacity for 3D object perception, limiting their suitability for applications requiring spatial depth and comprehensive scene understanding.

In computer vision, 3D object tracking is a well-established area [21, 22]. However, approaches to this task predominantly rely on specialized sensors, such as LiDAR, radar, or depth cameras, which diverge significantly from natural human perception. Humans typically rely on passive visual input (i.e., sight) and may also utilize supplementary cues like natural language descriptions (i.e., speech) to track objects without requiring active sensory input [34]. Inspired by this observation, this work introduces the task of **Mono**cular-Video-based **3D V**isual **L**anguage **T**racking (**Mono3DVLT**) - see Fig. 1.

The Mono3DVLT task focuses on 3D single object tracking from a monocular video guided by natural language cues. To facilitate research on this newly introduced task, we release a comprehensive dataset called Mono3DVLT-V2X, derived from V2X-Seq [48], comprising 79,158 segments of natural language descriptions that map to specific single object tracking within a monocular video. These descriptions are generated by ChatGPT [2] and then refined manually. This task bridges traditional machine object tracking and human-like object tracking by utilizing visual and language inputs to interpret real-world 3D objects through monocular videos.

We additionally develop an end-to-end neural network as the first tailored approach for the task, termed Mono3DVLT-MT, which is a 3D Visual Language Tracking Model utilizing a Memory-improved Token Turing Machine (TTM) [31] for Monocular Videos. Our network

[*] Equal Contribution. [†] Corresponding authors.

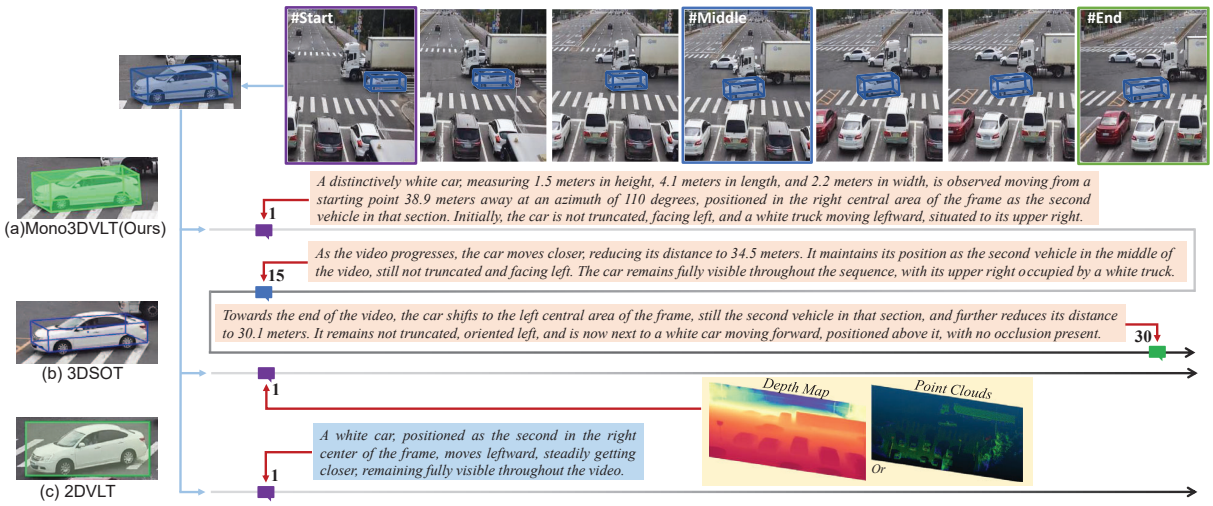

Figure 1. Monocular-Video-Based 3D Visual Language Tracking (Mono3DVLT) task. (a) Mono3DVLT aims to track the true 3D extent of referred object in a video using natural language descriptions. (b) Tracking specific objects is not feasible for monocular 3D objects tracking, as it needs depth map or point clouds [16, 21, 42]. (c) The counterpart of 2D VLT task [1, 19, 33] does not capture the 3D information of the referred object.

consists of a multi-modal feature extractor, a visual language encoder, a memory-augmented tracking decoder, and a tracking head - see § 4. We use RoBERTa [24] and the Swin Transformer [25] to extract language and visual features, respectively. A depth predictor is designed to explicitly learn geometry features. Next, to refine multi-scale visual and geometry features of the referred object, we propose a visual language encoder to perform language-guided feature learning based on pixel-wise attention. Subsequently, the memory-augmented tracking decoder refines the query. A learnable query gathers geometric features, language features, and visual features. Following each decoder step, critical data from the query is stored in the memory of TTM. The stored data aids in generating the query for the subsequent decoder iteration. In our model, visual-language-depth attention fuses object-level geometric cues and visual appearance into the query, fully enabling language-guided decoding.

Our main contributions are summarized as follows.

- We propose a new task, Mono3DVLT, for 3D visual language tracking in monocular RGB videos using language descriptions for state and appearance of a single object.
- We provide a large-scale dataset based on V2X-Seq, called Mono3DVLT-V2X, consisting of 79,158 detailed language descriptions for the proposed task.
- We design an end-to-end network, Mono3DVLT-MT, that fully integrates language, visual, and geometric features in a multi-modal framework, enhanced by a Memory-improved Token Turing Machine.
- We provide adequate benchmarks for Mono3DVLT. Extensive experiments show that our proposed Mono3DVLT-MT method significantly outperforms existing applicable methods.

2. Related Works

2.1. Single Object Tracking

Traditional single object tracking (SOT) is mainly based on the monocular video modality for feature extraction. In this field, Li et al. introduced SiamRPN++ [18], which demonstrated outstanding performance on multiple datasets, such as OTB2015 [43], VOT2018 [15], and LaSOT [9]. Similarly, Danelljan et al. proposed the ATOM tracker [7], which resulted in a performance boost on the TrackingNet dataset [27]. Furthermore, Bhat et al. leveraged scene context information to improve tracking accuracy [43] on the GOT-10k dataset [13]. These research efforts underscore the importance of embedding scene comprehension into tracking models.

2.2. 2D Visual Language Tracking

In recent years, the task of visual language tracking has predominantly focused on the 2D extent. Liu et al. proposed the Grounding DINO [23], forming the foundation for language-guided tracking tasks. Zuo et al. introduced a multiple object tracking framework [53] that leverages contrastive learning and modality alignment strategies to extract and fuse visual and language features, achieving state-of-the-art (SOTA) performance on the TNL2K dataset [40] and further pushing the boundary of visual language tracking. Furthermore, Li et al. proposed DTVLT [19], a comprehensive visual language tracking benchmark that uses

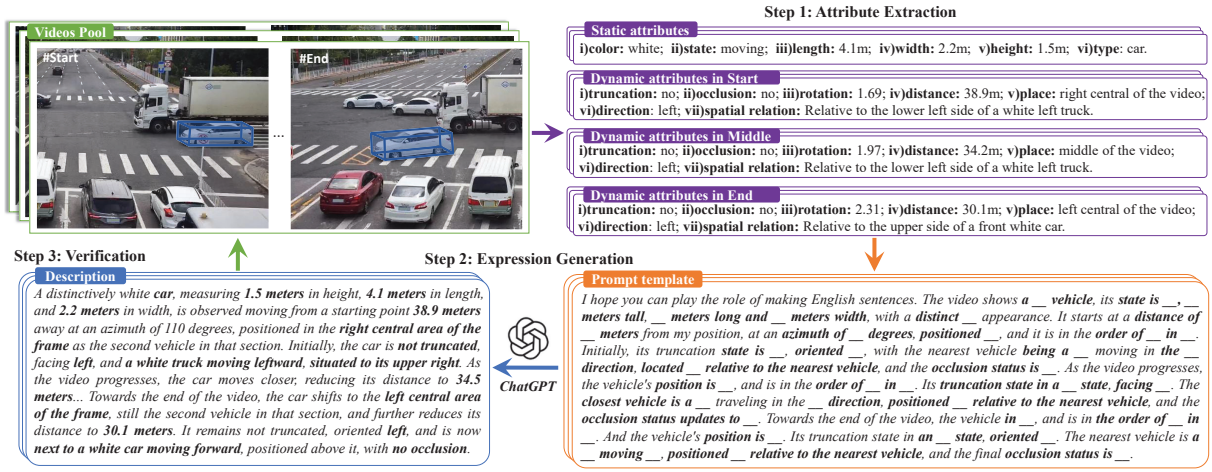

Figure 2. Our data generation pipeline: Step 1 - Global attributes and different temporal attributes that provide detailed information of the object are extracted. Step 2 - We fill in the prompt template designed with attributes, and input the complete prompt into ChatGPT to get descriptions. Step 3 - We verify that the description can uniquely identify the object.

Large Language Models (LLMs) to generate multi-modal tracking benchmarks, highlighting the potential of integrating language descriptions with visual cues for improved tracking accuracy across different modalities.

2.3. 3D Visual Language Tracking

Traditional 3D tracking methods mainly use multi-sensor data such as LiDAR or radar, which offer high-precision 3D data for object detection and tracking [41]. For instance, Zhou et al. proposed a Transformer-based 3D point cloud tracking model called PTTR [51], which achieves notable results on the Waymo Open Dataset [35]. Similarly, Chen et al. introduced VoxelNeXt [6], outperforming all previous LiDAR-based tracking methods on benchmarks such as nuScenes [3], Waymo [35], and Argoverse2. Yang et al. proposed MSMDFusion [14], which achieves fine-grained fusion between LiDAR and camera modalities to facilitate feature interaction. Language-guided 3D visual perception offers a promising direction by employing language information to enhance tracking performance. Zhan et al. proposed the Mono3DVG network [49], which fuses language features with visual features to achieve precise 3D object detection. Extending this idea to 3D object tracking leads to more accurate and reliable 3D visual language tracking in complex scenarios.

3. Mono3DVLT-V2X Dataset

We introduce Mono3DVLT, a new 3D Visual Language Tracking benchmark that leverages monocular video data. Unlike the previous 3D tracking paradigm that relies on expensive sensors, Mono3DVLT aims at utilizing monocular video data to make 3D tracking more accessible and practical for real-world applications. The proposed Mono3DVLT-V2X dataset, meticulously annotated with detailed language descriptions, provides a strong foundation for training and evaluating 3D VLT models. Using the large-scale V2X-Seq dataset [48], and rigorous validation, we ensure the high quality and reliability of the annotations.

Traditional datasets for Vision-Language tasks typically depend on manual annotation, which is costly and impractical for annotating large volumes of data. To this end, we extract attributes from the dataset and utilize ChatGPT to generate natural language descriptions, enabling the production of detailed descriptions. In Fig. 2, we illustrate the pipeline for generating language descriptions in Mono3DVLT-V2X by taking video frames and object attributes as inputs to produce concise and detailed descriptions for relevant objects, allowing large-scale text production at a lower cost through its three-stage process, discussed below.

Attribute extraction. We divide object attributes into static and dynamic categories. Static attributes include color (extracted using the EfficientNetV2 classification model [38]), state, length, width, height, and type. Dynamic attributes, on the other hand, encompass truncation, occlusion, rotation, distance, grid position, ordinal number, direction, and spatial relation. Notably, truncation, occlusion, length, width, and height are directly sourced from the labels of the raw V2X-Seq data. The state attribute of an object, whether it is "moving" or "parking", is derived from changes in its position over time. The distance and azimuth attributes are determined by the coordinates of the 3D bounding boxes, while the orientation, spatial relationship, grid position, and ordinal number attributes are based on 3D boxes and spatial geometry. The spatial relation attribute accounts for horizontal proximity (left, right), vertical proximity (upper,

Table 1. Comparison of Mono3DVLT-V2X dataset with existing tracking datasets.

Dataset	Publication	Visual Form	Language Form	Label	Task
VOT2018 [15]	ECCV'2018	RGB	-	2D bbox	2D Visual Tracking in RGB
TrackingNet [27]	ECCV'2018	RGB	-	2D bbox	2D Visual Tracking in RGB
H3D Dataset [28]	ICRA'2019	PC & RGB	-	3D bbox	3D Visual Tracking in Multi-sensor
nuScenes [3]	CVPR'2020	PC & RGB	-	2D/3D bbox	3D Visual Tracking in Multi-sensor
GOT-10k [13]	TPAMI'2021	RGB	-	2D bbox	2D Visual Tracking in RGB
Waymo Open Dataset [35]	ICCV'2021	PC & RGB	-	2D/3D bbox	3D Visual Tracking in Multi-sensor
LaSOT [9]	IJCV'2021	RGB	Manual	2D bbox	2D Visual Language Tracking in RGB
Mono3DVLT-V2X (Ours)	-	RGB	ChatGPT+Manual	2D/3D bbox	3D Visual Language Tracking in RGB

In the table, 'PC' denotes point cloud, 'Manual' refers to language annotations made by humans, and 'ChatGPT+Manual' means that the descriptions are generated by ChatGPT and then manually corrected.

Figure 3. The word cloud of language descriptions.

lower), and diagonal relations (lower left, upper right, etc.). To ensure the accuracy and correctness of these labeled attributes, we organized a team of five people to verify and correct the data.

Expression generation. We tailor the prompt template to produce expressions for ChatGPT. By inserting each attribute of the objects into the template and feeding the complete prompt to ChatGPT, we generate the descriptions. An associated word cloud of descriptions is provided in Fig. 3.

Verification. To guarantee the accuracy of descriptions, five people from our team jointly verify the dataset.

Using the established data generation pipeline, we present a comprehensive dataset Mono3DVLT-V2X, which consists of 79,158 video sequences accompanied by their corresponding natural language descriptions for SOT, based on the V2X-Seq dataset. Each description averages 176 words, giving deep insight into various attributes and behavior of objects, which enhances the performance of 3D object tracking systems. A comparative analysis of Mono3DVLT-V2X with other object tracking datasets is provided in Tab. 1.

4. Methodology

We introduce an end-to-end approach for the Mono3DVLT task introduced in this work. Our proposed network Mono3DVLT-MT serves as a strong baseline model for the Mono3DVLT-V2X dataset introduced above. This method is composed of four main components, as illustrated in Fig. 4, each of which will be elaborated upon in the subsequent sections.

4.1. Multi-modal Feature Extractor

We employ the pre-trained RoBERTa-base model [24] alongside a linear layer to extract language tokens denoted $\boldsymbol{f}_l^t \in \mathbb{R}^{C\times N_l^t}$, with N_l^t representing the sentence length. For each video frame $\boldsymbol{I}^t \in \mathbb{R}^{H\times W\times 3}$, the Swin Transformer [25] is employed with an additional linear layer to derive multi-scale visual features at four distinct levels, $\boldsymbol{f}_v^t \in \mathbb{R}^{C\times N_v^t}$, where $C = 256$ and N_v^t comprises $\frac{H}{8}\times\frac{W}{8} + \frac{H}{16}\times\frac{W}{16} + \frac{H}{32}\times\frac{W}{32} + \frac{H}{64}\times\frac{W}{64}$. To extract geometric features, we apply a lightweight depth predictor [50] to generate $\boldsymbol{f}_d^t \in \mathbb{R}^{C\times N_d^t}$, with $N_d^t = \frac{H}{16}\times\frac{W}{16}$.

4.2. Visual-Language Tracking Encoder

We develop a Visual Language encoder incorporating both visual and depth encoders. This setup enables inference of global contexts and creation of tokens with extended dependencies, represented as $\boldsymbol{f}_v^t \in \mathbb{R}^{C\times N_v^t}$ and $\boldsymbol{f}_d^t \in \mathbb{R}^{C\times N_d^t}$. As depicted in Fig. 5(a), our visual encoder architecture substitutes multi-head self-attention (MHSA) with multi-scale deformable attention (MSDA) to reduce the computational load associated with attention over multi-scale visual features. Additionally, we incorporate a multi-head cross-attention (MHCA) layer between the MSDA layer and the feed-forward network (FFN), which supplies language cues to the visual tokens.

To harness the appearance and geometry attributes of natural language descriptions, as depicted in Fig. 5 (b), our depth encoder is composed of a transformer encoder layer to process depth tokens. It employs the depth token $\boldsymbol{f}_d^t$ as a query for MHCA, using the language token $\boldsymbol{f}_l^t$ as both the key and the value. Subsequently, a multi-head attention (MHA) layer is deployed to apply implicit language-guided self-attention to geometry features, with original depth $\boldsymbol{f}_d^t$

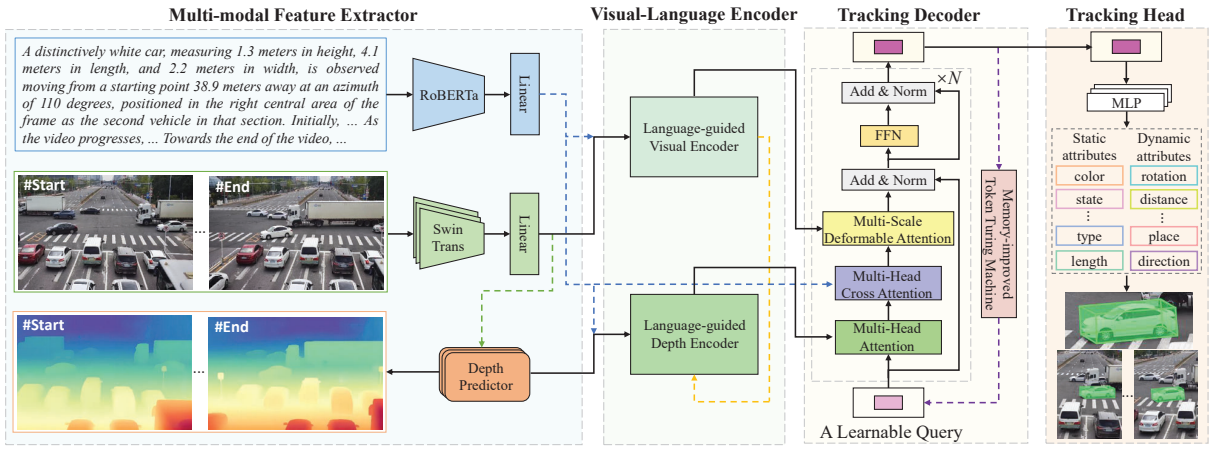

Figure 4. Overview of the proposed framework. The Multi-modal Feature Extractor first extracts textual, multi-scale visual, and geometry features. The Visual-Language Encoder refines visual and geometry features of referred objects based on pixel-wise attention. A learnable query fuses geometry cues and visual appearance of the object using visual-language-depth attention in the Tracking Decoder with Memory-improved Token Turing Machine. Finally, the Tracking Head employs multiple MLPs to predict the attributes of the object.

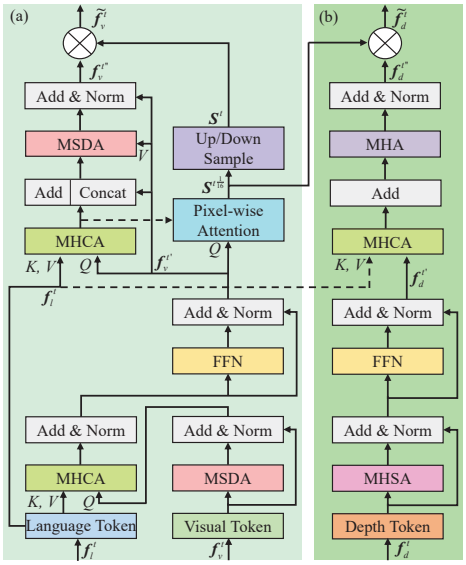

Figure 5. Details of Visual-Language Encoder. (a) Language-guided Visual Encoder. (b) Language-guided Depth Encoder.

serving as the value. The refined geometry feature is denoted as $\boldsymbol{f}_d^{t\,''}$.

The visual encoder, depicted in Fig. 5(a), requires the division and merging of multi-scale visual tokens $\boldsymbol{f}_v^t$ pre- and post-MHCA. This process employs $\boldsymbol{f}_v^{t\,'\frac{1}{16}}$—sized at $\frac{H}{16} \times \frac{W}{16}$—as the query. Subsequently, MSDA replaces MHA, resulting in the refined visual feature $\boldsymbol{f}_v^{t\,''}$.

We apply a linear projection to $\boldsymbol{f}_v^{t\,'\frac{1}{16}}$ along with the MHCA output from the visual tokens to reconstruct the original visual feature map $\boldsymbol{F}_{orig}^t \in \mathbb{R}^{C\times\frac{H}{16}\times\frac{W}{16}}$ and the language-associated feature map $\boldsymbol{F}_{lang}^t \in \mathbb{R}^{C\times\frac{H}{16}\times\frac{W}{16}}$. To investigate the alignment and detailed correlation between vision and language, we determine the attention score $\boldsymbol{s}_{ij}^t \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}}$ for each spatial region (i, j) in the feature map using the following formula:

$$\boldsymbol{F}_{orig}^t = \left\|\boldsymbol{F}_{orig}^t\right\|_2, \ \boldsymbol{F}_{lang}^t = \left\|\boldsymbol{F}_{lang}^t\right\|_2, \tag{1}$$

$$\boldsymbol{a}_{ij}^{t\,c} = \boldsymbol{F}_{orig}^{t\,c}(i,j) \odot \boldsymbol{F}_{lang}^{t\,c}(i,j), c = 1, 2, \ldots, C, \tag{2}$$

$$\boldsymbol{s}_{ij}^t = \sum_{c=1}^{C} \boldsymbol{a}_{ij}^{t\,c}, \tag{3}$$

where $\|\cdot\|_2$ signifies the ℓ_2-norm, and $\odot$ signifies the element-wise product.

Subsequently, we proceed to model the semantic similarity $\boldsymbol{S}^{t\,\frac{1}{16}}$, which entails dimensions of $\frac{H}{16} \times \frac{W}{16}$, for each pixel feature in relation to the text feature using a Gaussian function, as follows:

$$\boldsymbol{S}^{t\,\frac{1}{16}} = \alpha \cdot \exp\left(-\frac{(1-\boldsymbol{s}_{ij}^t)^2}{2\sigma^2}\right), \tag{4}$$

where α is the scaling factor and σ denotes the standard deviation, with both parameters being learnable.

Initially, we up-sample $\boldsymbol{S}^{t\,\frac{1}{16}}$ via bilinear interpolation, followed by downsampleing of $\boldsymbol{S}^{t\,\frac{1}{16}}$ through max pooling. The flattened score maps are then concatenated to derive the multi-scale attention score $\boldsymbol{S}^t \in \mathbb{R}^{N_v}$, as follows:

$$\boldsymbol{S}^t = \text{Concat}[\text{Up}(\boldsymbol{S}^{t\,\frac{1}{16}}), \boldsymbol{S}^{t\,\frac{1}{16}}, \text{Down}(\boldsymbol{S}^{t\,\frac{1}{16}}), \text{Down}(\boldsymbol{S}^{t\,\frac{1}{16}})]. \tag{5}$$

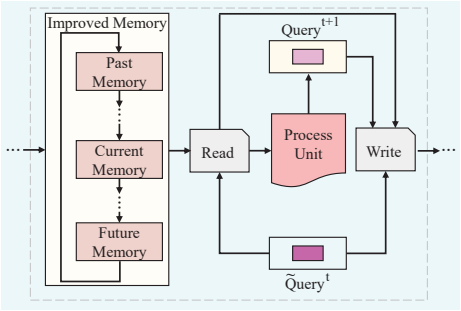

Figure 6. Details of Memory-improved Token Turing Machine.

Utilizing pixel-wise attention scores, the visual and geometric features are concentrated on regions relevant to the textual description. Employing the features $\boldsymbol{f}_v^{t\,''}$ and $\boldsymbol{f}_d^{t\,''}$, along with the scores (where $\boldsymbol{S}^{t\,\frac{1}{16}} \in \mathbb{R}^{N_d}$ is flattened), we execute element-wise multiplication to derive the adapted features of the referred object, as follows:

$$\tilde{\boldsymbol{f}}_v^t = \boldsymbol{f}_v^{t\,''} \cdot \boldsymbol{S}^t, \ \tilde{\boldsymbol{f}}_d^t = \boldsymbol{f}^{t\,''}_d \cdot \boldsymbol{S}^{t\,\frac{1}{16}}. \tag{6}$$

4.3. Memory-Augmented Tracking Decoder

In Fig. 4, the Tracking Decoder refines the query input by engaging memory. Specifically, the n^{th} decoder layer consists of a block composed of MHA, MHCA, and MSDA, and an FFN. A trainable query $\boldsymbol{Q}^t \in \mathbb{R}^{C\times 1}$ aggregates the initial geometric information, enhances this with text embedding to strengthen text-related geometric attributes, and finally assimilates appearance features from multi-scale visual data.

The Memory-improved Token Turing Machine (TTM) - details illustrated in Fig. 6, facilitates the connection between the output query of the decoder $\tilde{\boldsymbol{Q}}^t$ and the memory $\boldsymbol{M}$. Within this framework, memory is structured into blocks representing previous $\boldsymbol{M}^{t-1}$, current $\boldsymbol{M}^t$, and subsequent $\boldsymbol{M}^{t+1}$ states. The Read operation for the TTM is formally defined as follows:

$$\boldsymbol{P}^t = \text{Read}(\boldsymbol{M}^{t-1}, \boldsymbol{M}^t, \boldsymbol{M}^{t+1}, \tilde{\boldsymbol{Q}}^t). \tag{7}$$

This operation retrieves and integrates information from memory blocks in different temporal states (past, present, and future states). Reading and concatenating $\boldsymbol{M}^{t-1}$, $\boldsymbol{M}^t$, and $\boldsymbol{M}^{t+1}$, it produces the refined query $\boldsymbol{P}^t$. This process essentially transforms $\mathbb{R}^{3(m+n)\times d} \rightarrow \mathbb{R}^{3r\times d}$. The result is a query that incorporates information from multiple time steps, reducing the number of tokens passed to subsequent processing stages, and lowering the computational cost.

In the Memory-improved TTM, the processing unit updates the query to $\boldsymbol{Q}^{t+1} = \text{Process}(\boldsymbol{P}^t)$. We utilize a standard transformer [39] as the process units. Subsequently, the output $\boldsymbol{P}^{t+1}$, is written back to the memory employing the following equation:

$$\boldsymbol{M}^{t+1} = \text{Write}(\boldsymbol{M}^{t-1}, \boldsymbol{M}^t, \boldsymbol{M}^{t+1}, \tilde{\boldsymbol{Q}}^t, \boldsymbol{Q}^{t+1}). \tag{8}$$

This iterative mechanism refines and stores the query information, making it available for subsequent time steps, ensuring a continuous and efficient query refinement loop.

4.4. Tracking Head and Loss Function

The Tracking Head utilizes multiple MLPs to predict various attributes for each video frame. Specifically, the output of the decoder, denoted as $\tilde{\boldsymbol{Q}}^t \in \mathbb{R}^{C\times 1}$, represents learnable queries that are individually processed by linear layers. These layers predict the object category via a 3-layer MLP, estimate the 2D bounding box dimensions, and determine the 3D box center coordinates $(x_{3D}, y_{3D}, l, r, t, b)$. Here, (l, r, t, b) denotes the left, right, top and bottom of the 2D bounding box. A separate 2-layer MLP is used to predict the 3D box dimensions, orientation, and depth $(h_{3D}, w_{3D}, l_{3D}, \theta, d_{reg})$. The predicted depth, d_{pred}, is computed following the method in Zhang et al. [50]. The attributes predicted for each frame are then aggregated to provide a coherent interpretation of the video sequence.

In the subsequent process, the 2D loss function $\mathcal{L}_{2D}$ comprises four components: $\mathcal{L}_{class}$ represents the focal loss [20] for various classes, $\mathcal{L}_{lrtb}$ is the L1 loss for the four sides of the 2D bounding box, $\mathcal{L}_{xy3D}$ is the L1 loss for the 3D center, and $\mathcal{L}_{GIoU}$ is the GIoU loss [29] used to refine the 2D bounding box. The 2D loss function is defined as:

$$\mathcal{L}_{2D} = \lambda_1\mathcal{L}_{class} + \lambda_2\mathcal{L}_{lrtb} + \lambda_3\mathcal{L}_{GIoU} + \lambda_4\mathcal{L}_{xy3D}, \tag{9}$$

where the coefficients $\lambda_{1\sim 4}$ are set to $(2, 5, 2, 10)$ following Zhang et al. [50]. Additionally, the Hungarian algorithm [17] is employed for optimal query-to-Ground-Truth matching based on the 2D loss function $\mathcal{L}_{2D}$.

Simultaneously, the 3D loss function $\mathcal{L}_{3D}$ includes three components: $\mathcal{L}_{size3D}$, the 3D IoU-oriented loss [29] for evaluating 3D dimensions; $\mathcal{L}_{orien}$, the Multi-Bin loss [4] to estimate orientation; and $\mathcal{L}_{depth}$, the Laplacian algebraic uncertainty loss [5] for depth estimation. The 3D loss function is formulated as:

$$\mathcal{L}_{3D} = \mathcal{L}_{size3D} + \mathcal{L}_{orien} + \mathcal{L}_{depth}. \tag{10}$$

Furthermore, $\mathcal{L}_{dmap}$, which uses the focal loss for the predicted depth map categories, is incorporated into the cumulative losses. The final loss equation is as follows:

$$\mathcal{L}_{overall} = \mathcal{L}_{2D} + \mathcal{L}_{3D} + \mathcal{L}_{dmap}. \tag{11}$$

5. Experiments

5.1. Experimental Setup

The proposed Mono3DVLT-V2X dataset includes a total of 56,106 video sequences paired with linguistic annotations

Table 2. Comparison of Mono3DVLT-MT with baselines, the bold means best performance, and the underline represents the second best.

Method	SR@0.5 (↑)	SR@0.6 (↑)	SR@0.7 (↑)	SR@0.8 (↑)	SR@0.9 (↑)	AOR (↑)	PR@1.0 (↑)	ACE (↓)
ZSGNet + backproj [32]	37.29	30.93	25.37	21.34	19.27	40.69	37.41	1.132
FAOA + backproj [45]	35.43	31.55	29.62	23.53	18.69	41.18	43.40	1.059
ReSC + backproj [46]	50.33	46.95	38.41	32.56	27.36	53.43	68.52	0.792
TransVG + backproj [8]	54.50	47.93	41.62	37.63	23.83	58.62	66.78	0.783
JointNLT + backproj [52]	61.40	57.33	51.70	49.03	45.27	68.31	70.43	0.697
Mono3DVG-TR [49]	71.75	67.85	63.47	56.26	49.13	79.13	75.89	0.594
Mono3DVLT-MT (Ours)	**81.63**(+9.88)	**75.41**(+7.56)	**68.94**(+5.47)	**62.63**(+6.37)	**58.86**(+9.93)	**85.12**(+5.99)	**81.56**(+5.67)	**0.521**(-0.073)

for training purposes, 11,203 sequences for validation and an additional 11,849 for testing. The model employs the AdamW optimizer, with the learning rate and weight decay parameters configured to 1×10^{-4}. Compared to other existing methods adapted to our task in this dataset, our Mono3DVLT-MT demonstrates outstanding performance across comprehensive metrics.

5.2. Evaluation Metrics

We employ Success Rate (SR) [44] at Intersection over Union (IoU) [30] thresholds as the core metric. In addition, Average Overlap Rate (AOR) [44] for 3D bounding box alignment, Precision Rate (PR) [26] calculated via Euclidean distance between predicted and ground-truth centers, and Average Center Error (ACE) [26] quantifying positional offset magnitude.

5.3. Baselines

For the proposed Mono3DVLT task, we designed a set of benchmark experiments and implemented a standardized methodology to evaluate the effectiveness of different approaches. We included multiple text-guided 2D grounding and VLT methods such as ZSGNet [32], FAOA [45], ReSC [46], TransVG [8], and JointNLT [52]. These methods, when integrated with inverse projection and object tracking methods, facilitate efficient 3D tracking. By comparing these baselines with our proposed model, we can systematically assess the performance of each model.

5.4. Results Analysis and Visualization

We perform a comparative analysis of the Mono3DVLT-MT model against alternative frameworks to evaluate its capabilities. This evaluation is based on a standardized framework, with detailed findings presented in Tab. 2. In particular, our Mono3DVLT-MT model outperformed the other models, attaining an average overlap rate (AOR) of 85.12%. This marks an improvement of approximately 5.99% over the VLT model. Furthermore, Mono3DVLT demonstrated remarkable performance in all specified thresholds (SR@0.5-0.9). The model shows excellent accuracy in predicting 3D bounding boxes, achieving an Average Center Error (ACE) as low as 0.521 pixels. The predicted bounding boxes are aligned well with the size and orientation of the ground truth boxes, evidencing superior center prediction accuracy over various distance thresholds.

Table 3. Ablation study for the proposed components of our approach. 'Res.' and 'Swin.' denote ResNet50 and Swin Transformer, 'M.TTM' denotes Memory-improved TTM. The '✓' indicates that the component is used, while a blank space indicates that the component is not used. The bold means best performance, and the underline represents the second best.

Extractor		Decoder	Mertic			
Res.	Swin.	M.TTM	SR@0.9(↑)	AOR(↑)	PR@1.0(↑)	ACE(↓)
✓			49.13	79.13	75.89	0.594
✓		✓	53.91	80.28	81.03	0.559
	✓		55.38	84.32	80.68	0.589
	✓	✓	**58.86**	**85.12**	**81.56**	**0.521**

Visual comparisons (see Fig. 7) highlight Mono3DVLT-MT's superior center prediction accuracy and high IoU scores. In contrast, TranVG backproj struggles with center misalignment, while Mono3DVG improves center accuracy but still faces size/orientation errors.

On the other hand, TranVG backproj is based exclusively on single-image modalities and text-guided features for tracking, with 3D box predictions derived from 2D box back projections. This reliance can impair accuracy due to the inherent dependence on 2D box precision, potentially yielding suboptimal outcomes. In contrast, adeptly integrating visual, linguistic, and geometric information is a pivotal strength of our method for achieving precise 3D object tracking in monocular videos.

5.5. Ablation Study

To validate the effectiveness of the proposed Mono3DVLT-MT components, we performed detailed ablation studies on the feature extractor and the memory-improved TTM of the tracking decoder. These experiments demonstrate the effectiveness of these components within Mono3DVLT-MT.

Visual Feature Extractor. This module is responsible for extracting features from each frame of the video, which are essential for queries analyzing 3D tracking cues within the cross-attention layer. To evaluate the effectiveness of

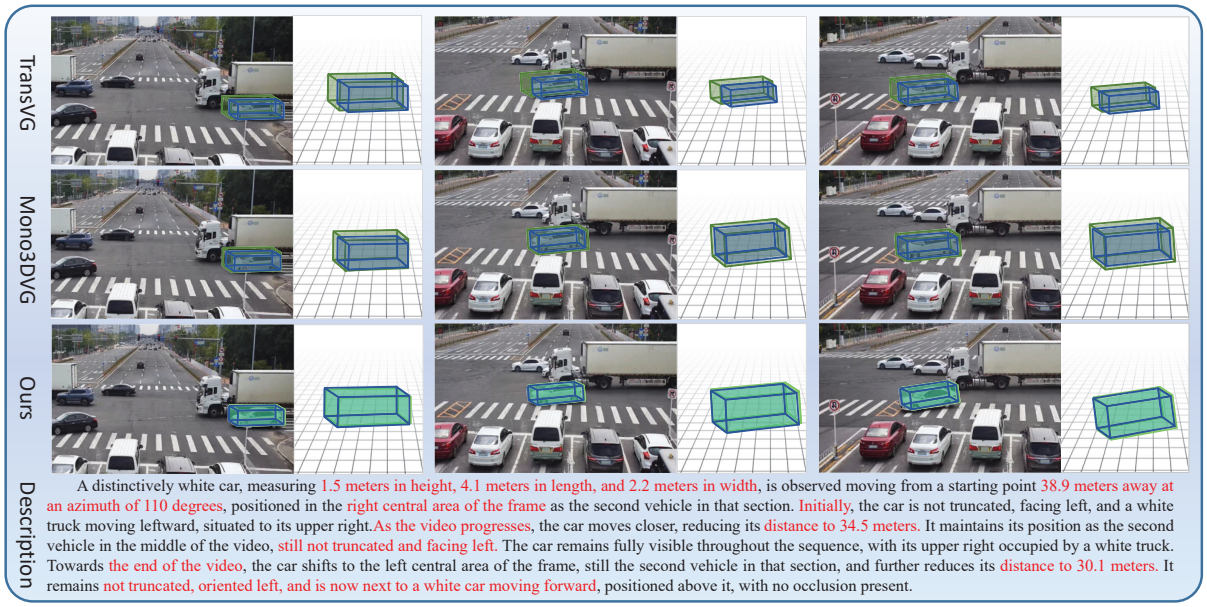

Figure 7. Visualization of results of our model alongside the best performing existing models Predicted and ground truth 3D bounding boxes are also shown separately. 'TransVG' refers to obtaining 2D bounding boxes guided by semantics through TransVG and then projecting them back to 3D bounding boxes. Mono3DVG and our model directly obtain 3D bounding boxes. Blue boxs denote the ground truth, and green boxes denote the predictions.

the feature extractor, we performed ablation studies comparing ResNet50 and Swin Transformer. The experimental results, as shown in Tab. 3, demonstrate that the Swin feature extractor significantly outperforms ResNet50. We attribute this superiority to the fact that the Swin Transformer, pre-trained on large datasets, has learned to capture more nuanced features. Additionally, its sliding-window mechanism and cross-window connection allow for more effective feature extraction from each frame.

Memory-Augmented Tracking Decoder. We conduct ablation studies to investigate the contribution of the Memory-Augmented Tracking Decoder. As shown in Tab. 3, the Memory-Augmented Tracking Decoder outperforms the standard Decoder without memory augmentation. This improvement can be attributed to the Memory-improved TTM, which extracts relevant features from the previous frame's decoder and stores them in memory. This allows the model to retain information about objects that match the language description in historical frames. As a result, during the next frame, the query can retrieve relevant features from memory, facilitating an iterative process that continuously optimizes and stores query information for efficient tracking in subsequent time steps.

6. Conclusion

We propose a framework to track objects in video sequences with additional textual cues. To the best of our knowledge, our work is the first attempt towards 3D visual-language tracking using monocular video. Towards that end we first propose Mono3DVLT task to better emulate human tracking ability using passive visual and language cues. We then build a large dataset to create a challenging experimental environment for the proposed task. Finally, we designed an end-to-end trainable neual network tailored to the proposed task. With extensive experiments on the proposed dataset, we demonstrate that our method significantly outperforms the benchmark methods that can be adapted for the task. We hope that this contribution opens up a new avenue in monocular video based 3D object tracking.

Acknowledgements

This work was supported by the National Key R&D Program of China (Grant No. 2023YFB4301800). Ajmal Mian is the recipient of an Australian Research Council Future Fellowship Award (project number FT210100268) funded by the Australian Government. This work was also supported by the Fundamental Research Funds for the Central Universities, CHD (Grant No. 300102244202), and National Natural Fund Joint Fund Project (Grant No.U21B2041).

References

[1] Adnen Abdessaied, Lei Shi, and Andreas Bulling. Multimodal video dialog state tracking in the wild. In *Computer Vision – ECCV 2024*, pages 348–365, Cham, 2025. Springer Nature Switzerland. 2

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 3, 4

[4] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision – ECCV 2020,Lecture Notes in Computer Science*, page 202–221, 2020. 6

[5] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[6] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[7] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[8] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 7

[9] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4

[10] Heng Fan, Longyin Wen, Dawei Du, Pengfei Zhu, Qinghua Hu, Haibin Ling, Mubarak Shah, Biao Wang, Bin Dong, Di Yuan, et al. Visdrone-sot2020: The vision meets drone single object tracking challenge results. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 728–749. Springer, 2020. 1

[11] Shiyu Hu, Dailing Zhang, wu meiqi, Xiaokun Feng, Xuchen Li, Xin Zhao, and Kaiqi Huang. A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship. In *Advances in Neural Information Processing Systems*, pages 25007–25030. Curran Associates, Inc., 2023. 1

[12] Shiyu Hu, Xin Zhao, Lianghua Huang, and Kaiqi Huang. Global instance tracking: Locating target more like humans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):576–592, 2023. 1

[13] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2021. 2, 4

[14] Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21643–21652, 2023. 3

[15] Matej Kristan, Ales Leonardis, Jiri Matas, Michael Felsberg, Roman Pflugfelder, Luka ˇCehovin Zajc, Tomas Vojir, Goutam Bhat, Alan Lukezic, Abdelrahman Eldesokey, et al. The sixth visual object tracking vot2018 challenge results. In *Proceedings of the European Conference on Computer Vision (ECCV) workshops*, pages 0–0, 2018. 2, 4

[16] Sheng-Yao Kuan, Jen-Hao Cheng, Hsiang-Wei Huang, Wenhao Chai, Cheng-Yen Yang, Hugo Latapie, Gaowen Liu, Bing-Fei Wu, and Jenq-Neng Hwang. Boosting online 3d multi-object tracking through camera-radar cross check. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 2125–2132, 2024. 2

[17] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[18] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[19] Xuchen Li, Xiaokun Feng, Shiyu Hu, Meiqi Wu, Dailing Zhang, Jing Zhang, and Kaiqi Huang. Dtllm-vlt: Diverse text generation for visual language tracking based on llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7283–7292, 2024. 1, 2

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 6

[21] Jiaming Liu, Yue Wu, Maoguo Gong, Qiguang Miao, Wenping Ma, Cai Xu, and Can Qin. M3sot: Multi-frame, multi-field, multi-space 3d single object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3630–3638, 2024. 1, 2

[22] LiChen Liu, XiangYu Song, HuanSheng Song, ShiJie Sun, Xian-Feng Han, Naveed Akhtar, and Ajmal Mian. Yolo-3dmm for simultaneous multiple object detection and tracking in traffic scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 25(8):9467–9481, 2024. 1

[23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2, 4

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 4

[26] Wafae Mrabti, Kaoutar Baibai, Benaissa Bellach, Rachid Oulad Haj Thami, and Hamid Tairi. Predicting learning tracking: A comparative study. In *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*, pages 118–123, 2018. 7

[27] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 4

[28] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In *International Conference on Robotics and Automation*, 2019. 4

[29] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[30] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, 2019. 7

[31] MichaelS. Ryoo, Keerthana Gopalakrishnan, Kumara Kahatapitiya, Ted Xiao, Kanishka Rao, Austin Stone, Yao Lu, Julian Ibarz, and Anurag Arnab. Token turing machines. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19070–19081, 2022. 1

[32] Arka Sadhu, Kan Chen, and Ram Nevatia. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4694–4703, 2019. 7

[33] Yanyan Shao, Shuting He, Qi Ye, Yuchao Feng, Wenhan Luo, and Jiming Chen. Context-aware integration of language and visual references for natural language tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19208–19217, 2024. 2

[34] Adva Shoham, Idan Daniel Grosbard, Or Patashnik, Daniel Cohen-Or, and Galit Yovel. Using deep neural networks to disentangle visual and semantic information in human perception and memory. *Nature Human Behaviour*, pages 1–16, 2024. 1

[35] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 4

[36] Shijie Sun, Naveed Akhtar, Xiangyu Song, Huansheng Song, Ajmal Mian, and Mubarak Shah. Simultaneous detection and tracking with motion modelling for multiple object tracking. In *Computer Vision – ECCV 2020*, pages 626–643, Cham, 2020. Springer International Publishing. 1

[37] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):104–119, 2021. 1

[38] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021. 3

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanN. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neural Information Processing Systems,Neural Information Processing Systems*, 2017. 6

[40] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13763–13773, 2021. 2

[41] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3d multi-object tracking: A baseline and new evaluation metrics, 2020. 3

[42] Qiao Wu, Kun Sun, Pei An, Mathieu Salzmann, Yanning Zhang, and Jiaqi Yang. 3d single-object tracking in point clouds with high temporal variation. In *Computer Vision – ECCV 2024*, pages 279–296, Cham, 2025. Springer Nature Switzerland. 1, 2

[43] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015. 2

[44] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12549–12556, 2020. 7

[45] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019. 7

[46] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404. Springer, 2020. 7

[47] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4):13–es, 2006. 1

[48] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5486–5495, 2023. 1, 3

[49] Yang Zhan, Yuan Yuan, and Zhitong Xiong. Mono3dvg: 3d visual grounding in monocular images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6988–6996, 2024. 3, 7

[50] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Xuanzhuo Xu, Yu Qiao, Peng Gao, and Hongsheng Li. Monodetr: depth-guided transformer for monocular 3d object detection. *arXiv preprint arXiv:2203.13310*, 2022. 4, 6

[51] Changqing Zhou, Zhipeng Luo, Yueru Luo, Tianrui Liu, Liang Pan, Zhongang Cai, Haiyu Zhao, and Shijian Lu. Pttr: Relational 3d point cloud object tracking with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[52] Li Zhou, Zikun Zhou, Kaige Mao, and Zhenyu He. Joint visual grounding and tracking with natural language specification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23151–23160, 2023. 7

[53] Jixiang Zuo, Tao Wu, Meiping Shi, Xueyan Liu, and Xijun Zhao. Multi-modal object tracking with vision-language adaptive fusion and alignment. In *2023 5th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI)*, pages 1125–1133, 2023. 2