Using Multi-cues Fusion in Language-Guided Multiple Object Tracking

Yangkai Chen , Shipeng Hong , Fan Xiao , Yixin Li

23020241154377,36920241153210,36920241153260,36920241153229
Yangkai Chen from Information Institute class,rest of all from AI class

Abstract

Multi-object tracking is a fundamental task in the field of computer vision, aiming to predict the trajectories of objects in video sequences. The Refering MOT task, building upon traditional MOT, uses natural language to guide models in predicting trajectories solely for the objects of interest, leveraging the complementary information between language and vision modalities for more precise guidance. Previous approaches integrated text modules into trackers, typically using text features to guide the generation of visual features. However, these methods often overlook other clues during the tracking process, such as motion and appearance cues, which also play crucial roles in multi-object tracking, particularly in complex tracking scenarios. In this study, we introduce a multi-cue fusion mechanism into the process of language-guided visual feature generation. Unlike earlier MOT methods, we employ an early fusion approach to improve performance in the association stage, avoiding complex heuristic post-processing methods. Experimental results demonstrate that our approach, using a simple fusion method, outperforms previous state-of-the-art methods on the Refering-KITTI dataset, achieving 46.988 and 14.514 for HOTA and MOTA metrics, respectively.

Introduction

In the realm of computer vision, multiple object tracking stands as a foundational task of paramount importance. It is designed to concurrently handle identity confirmation and tracking tasks within videos. The practical applica-tions of MOT, such as in autonomous driving, surveillance, and video analysis, underscore its critical role in everyday scenarios. Presently, the core of most MOT models lies in the extraction of visual features for object localization and tracking. While early tracking methodologies relied on manually crafted visual features to delineate distinct iden-tity information, contemporary approaches pivot towards utilizing deep neural networks for discerning visual features to discriminate identities. However, the exclusive reliance on visual features for identity discrimination and tracking grapples with pronounced challenges in intricate settings marked by dense occlusions and motion blur.

To elevate the precision of MOT models, the prevalent methodology entails the utilization of high-precision object detection models to extract Region of Interest (RoI) information within each frame. Thereafter, a visual model is engaged to glean visual features for individual RoIs, facilitating object associations crucial for tracking. The linchpin for achieving high-precision tracking, assuming the object detection model captures all targets necessitating tracking, lies in the extraction of robust visual features and the enactment of precise association processes. Nevertheless, this approach falls short in addressing the challenges that MOT encounters in complex scenarios.

Propelled by the strides in multimodal technologies, numerous studies have homed in on integrating language modality information into visual tasks, showcasing promising capabilities and bolstered generalization abilities. Notably, in select literature(Awais et al. 2023; Du et al. 2022; Srivatsan, Naseer, and Nandakumar 2023), language descriptions have been validated for furnishing supplementary information to the visual modality during the association phase, there-by fortifying the feature set for multiple object tracking.

Previous works have simply integrated textual modules into existing trackers by taking visual features extracted from the backbone as input and generating new visual features guided by language information. This simplistic integration approach only yields marginal improvements, overlooking many crucial tracking cues. Consequently, even with language-guided assistance, negative performance outcomes may arise, especially in complex scenarios. Traditional multi-object tracking algorithms commonly employ multi-cue fusion techniques to enhance tracking performance, often fusing semantic, appearance, and motion cues. However, in language-guided scenarios, typically only semantic cues are utilized while appearance and motion cues are disregarded. Yet, these cues can offer richer information to enhance tracking performance throughout the tracking process.

The synergistic interaction between semantic and motion cues is evident due to the strong correlation between motion patterns and category information. Instances corresponding to semantically similar categories typically exhibit similar motion patterns. If a model learns a certain motion pattern during training, this knowledge can be directly generalized to semantically related categories to address complex scenarios. Appearance cues are widely considered a nec-

essary condition for precise tracking(Li et al. 2023). Our approach integrates semantic, appearance, and motion cues through fusion embedding into the matching process. This integration method starkly contrasts with traditional multi-cue fusion methods, as conventional approaches often rely on heuristic-based fusion in later stages.

We conducted extensive experiments on the Refer-KITTI dataset to evaluate our approach. The results demonstrate that our method outperforms previous solutions.Specifically, our method achieves a 5.4% higher HOTA, 49.8% higher MOTA, and 6.4% higher IDF1 compared to the previous state-of-the-art method iKUN(Du et al. 2024).

Related work

Multi Object Tracking

In the realm of multi-object tracking algorithms, the predominant paradigm is tracking-by-detection. In this framework, objects in each frame are initially identified by an object detection model, followed by an association al-gorithm that correlates objects and trajectories across frames. Consequently, the majority of algorithms concen-trate on the association phase. SORT(Bewley et al. 2016) uti-lizes Kalman filtering for motion modeling and employs Intersection over Union (IoU) for association. Expanding upon SORT, DeepSORT(Wojke, Bewley, and Paulus 2017) integrates convolutional neural networks to extract appearance features of objects. Additionally, ByteTrack(Zhang et al. 2022) and Strong-SORT(Du et al. 2023) extend more robust association rules and post-processing strategies. Evidently, these algo-rithms rely solely on a single visual modality for associat-ing objects and trajectories. Their performance sharply de-clines when handling complex scenarios like occlusions and motion blur.

Due to the powerful modeling capabilities of the Transformer architecture, an increasing number of end-to-end trackers are being built using the Transformer architecture. These models(Zeng et al. 2022; Zhang, Wang, and Zhang 2023; Yu et al. 2023) utilize the interaction between object information and global image information within the transformer modules to obtain more comprehensive correlations, re-sulting in enhanced performance. However, trackers based on the Transformer architecture still exhibit relatively poor target localization and tracking performance in dense scenes.

Refering Tracking

In recent years, single-object tracking has garnered significant attention, with an increasing number of studies incorporating the language modality into single-object tracking algorithms. By leveraging the complementary relationship between language and visual modalities, these algorithms aim to enhance tracking robustness, most of which are based on Transformer architectures. JointNLT(Zhou et al. 2023) directly embeds language descriptions, templates, and textual images into a Transformer encoder for associ-ation modeling. OVLM(Zhang et al. 2023) introduces a memory token selection mechanism to filter redundant to-kens using textual information. (Botach, Zheltonozhskii, and Baskin 2022) proposes a multimodal module that decodes instance-level features into different modal sequences for tracking. MMTrack(Zheng et al. 2023) converts language descriptions and bounding boxes into discrete tokens, transforming the referring track task into a token generation task. iKUN(Du et al. 2024) presents a knowledge fusion module that integrates information from text and image streams to guide the generation of more robust visual fea-tures for tracking.

In recent time, an increasing number of MOT approach-es are referencing the practices of single-object tracking, utilizing language modalities to guide the generation of visual features for tracking objects of interest. OVTrack(Li et al. 2023) leverages features generated by CLIP's image and text encoders as supervisory signals to guide the tracking model in producing similar visual and textual features, thereby achieving more robust tracking performance. LaMOT(Li et al. 2024a) introduces a paradigm where visual and language modalities mutually guide the selection of features for feature enhancement. TransRMOT(Wu et al. 2023) integrates language modalities for simple fusion based on the end-to-end mod-el MOTR(Zeng et al. 2022). TempRMOT(Zhang et al. 2024) treats all targets as queries, merging visual and textual features through a cross-attention mechanism to track queries of interest based on the fusion results. LG-MOT(Li et al. 2024b) employs graph neural networks for object association, uti-lizing multi-granularity language descriptions to align lan-guage features at different granularities with node and edge embeddings of the graph, optimizing the association effect. Due to the complementary relationship between language and visual modalities, these algorithms have all shown promising results.

Method

We first provide an overview(figure 1) of the method we propose, followed by detailed explanations on how we acquire semantic, motion, and appearance cues and then integrate them. Our method is based on the iKUN(Du et al. 2024) architecture, a two-stage tracker. In the first stage, the detector provides RoI regions as candidate objects and trajectories. In the second stage, the CLIP(Radford et al. 2021) architecture is used to extract visual features for each RoI region. Simultaneously, text information is employed to generate text embeddings as guiding information to produce visual embeddings. These visual embeddings are then utilized for association matching. The fusion of multiple cues occurs in the second stage, before the generation of visual embeddings guided by language. Semantic, appearance, and motion cues are generated by corresponding output head and participate in the subsequent language-guided process.

Semantic Head

In the iKUN architecture, visual semantic information is naturally extracted. In the original iKUN model, the CLIP image encoder is used to extract visual features for each RoI. Thanks to the powerful image-text matching capability of CLIP, it can be assumed that the extracted visual features contain the semantic information of each RoI. However, due

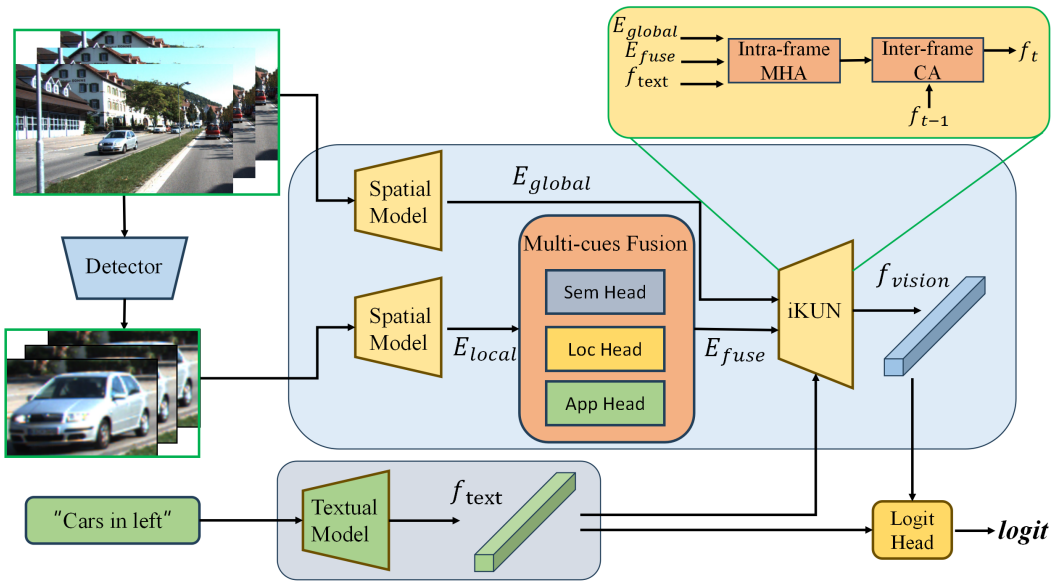

Figure 1: **Pipeline of our method:**The detector takes input images and outputs candidate RoI. Then the RoI regions and the global image input extract global embedding E_{global} and local embedding E_{local} through the visual flow. E_{local} is fused into a fused embedding E_{fuse} by the Multi-Cues Fusion module, incorporating information from multiple cues. E_{global} and E_{fuse} pass through the iKUN architecture, while the Textual Model extracts textual prompts to obtain textual feature f_{text} inputted into iKUN for language guidance, resulting in guided visual feature f_{vision}. In the diagram, MHA represents Multi-Head Attention mechanism, and CA represents Cross Attention mechanism.

to the complexity of the CLIP model, directly using the CLIP encoder to extract semantic cues can lead to high computational costs. Therefore, we adopt knowledge distillation to obtain semantic cues.Specifically, we input the proposals provided by the detector into a simple convolutional neural network for adaptive classification. The classification head is distilled by the CLIP image encoder to output semantic embeddings aligned with CLIP. Subsequently, we add a five-layer MLP to map the semantic embeddings and obtain the final semantic cues, denoted as E_{sem}.

Location Head

The Location Head is responsible for assisting in extracting motion cues. In traditional multi-object tracking, motion cues rely on Kalman filtering under the linear motion assumption. However, such an assumption is not always valid in all scenarios. Therefore, our work avoids using Kalman filtering as the sole method for extracting motion cues. Considering that motion cues contain information such as the instance's motion speed, direction, and acceleration, for computational convenience, we divide the acquisition of motion cues into two parts. The first part involves obtaining positional embeddings through the instance's location information, while the second part utilizes attention mechanisms both inter-frame and intra-frame to interact and obtain the final motion cues. Such cues do not rely on the linear motion assumption but instead derive motion information from the instance's positional changes. In this section, we only discuss the acquisition of positional embeddings.

The Location Head receives bounding boxes from the detector as input. To ensure the stability and accuracy of model training, it is necessary to scale all bounding boxes to a unified scale. For a given image of size [H, W], for each bounding box (x_1, y_1, x_2, y_2), we compute the normalized bounding box coordinates based on these two dimensions. The process involves first shifting the original coordinates to relative coordinates with respect to the image center. We obtain the image center coordinates as (C_x, C_y), where $C_x = W/2$ and $C_y = H/2$. To ensure scale consistency, we additionally introduce a scaling factor, typically set as $s = 0.7 \cdot max(H, W)$. The normalized coordinates are then expressed as follows:

$$(x', y', w', h') = (\frac{x_1 - C_x}{s}, \frac{y_1 - C_y}{s}, \frac{x_2 - x_1}{s}, \frac{y_2 - y_1}{s})$$

The normalized coordinates are input into the Location Head to obtain the positional embedding E_{loc}.In our work, we have constructed a simple five-layer MLP, which shares a similar architecture with the MLP used in the Semantic Head, differing only in the input dimensions.

Appearance Head

The Appearance Head is utilized for extracting appearance cues. In our work, we employ ResNet50 as the backbone network for appearance cue extraction. Each RoI provided by the detector is used as input, and the output serves as visual embeddings optimized for the association phase. Apart from utilizing the backbone network to extract visual embeddings, the appearance head includes an additional fully

connected layer. This layer maps the visual embeddings to the appearance cue space, yielding appearance cues E_{app} consistent with other cue dimensions.

Multi-cues Fusion

Before proceeding with language guidance, we represent each object in a frame with rich contextual cues rather than relying on singular feature representations. Hence, the three cues mentioned earlier, semantic cues E_{sem}, positional embedding E_{loc}, and appearance cues E_{app}, are fused to obtain an initial visual representation. These embeddings encapsulate various aspects of an instance's information, enabling more precise matching in subsequent language guidance and alignment processes by leveraging not only category information but also contextual information from appearance features and spatial positions. For convenience, we need to merge the extracted cue information into a unified space. Due to the unique nature of motion cues, we must preserve the original shape of positional cues. Therefore, our fusion process employs a simple yet effective additive operation as follows:

$$E^i_{fus} = E^i_{sem} + E^i_{loc} + E^i_{app}$$

Here, E^i_{fus} represents the fused feature corresponding to the i-th instance in the frame. This approach ensures the incorporation of all three cues and, due to the nature of the additive operation, allows positional information to directly participate in subsequent intra-frame and inter-frame attention mechanism calculations.

Inter-Frame Cross-Attention

In the original iKUN architecture, the language-guided process is achieved through three different multimodal feature fusion mechanisms. By taking global features, local features, and textual features of images as input, the architecture generates new visual features. Among the three fusion mechanisms proposed in the original text, both global features and local features undergo computations using self-attention mechanisms, which can be seen as facilitating intra-frame information interaction. However, the iKUN architecture lacks an inter-frame attention mechanism, opting instead for a simple average pooling method to aggregate temporal information between frames. This approach is considered crude. In our work, we employ an inter-frame cross-attention mechanism to extract temporal information between frames.Specifically, after intra-frame interactions, the model utilizes cross-attention mechanisms to simultaneously process reference frames and key frames, enabling the model to capture richer temporal information. This practice is common in the field of multi-object tracking. The specific calculations involved are as follows:

$$CA_{KR}\left(Q_K, K_R, V_R\right) = \sigma\left(\frac{Q_K K_R^T}{\sqrt{d}}\right) V_R$$

$$CA_{RK}\left(Q_R, K_K, V_K\right) = \sigma\left(\frac{Q_R K_K^T}{\sqrt{d}}\right) V_K$$

Where Q_K and Q_R are query vectors from the key frame and reference frame respectively, K_R and K_K are key vectors from the reference frame and key frame respectively, and V_R and V_K are value vectors from the reference frame and key frame respectively. The symbol σ denotes the softmax operation. This step is crucial for the model's ability to stably track objects over time, enabling the model to comprehend the most relevant features of objects during temporal changes.

Experiment

Evaluation Metrics

Following existing MOT models, we employ IDF1,HOTA, MOTA,and IDSW as metrics for evaluating model performance. IDF1 prioritizes the duration of tracking a specific object, primarily assessing tracking continuity and reidentification accuracy.HOTA measures the accuracy of detection and association. IDSW represents the total numberof identity switches, while MOTA focuses more on evaluating detection accuracy.

Benchmark

We evaluated our approach on the Refer-KITTI dataset. Refer-KITTI is currently the only dataset designed for refering multi-object tracking. This dataset is an extension of the KITTI dataset, where each video sequence in KITTI is associated with one or more textual descriptions. We utilized 15 videos with 80 distinct descriptions for training and 3 videos with 63 different descriptions for testing. Our method was assessed using validation metrics on this dataset.In comparison, we use iKUN as our baseline model, and additionally, our method is compared with some traditional trackers and Refering MOT trackers.

Implementation Details

In our experiments, we adopted YOLOv8 as our detector. For the language-guided architecture, we employed the CLIP-RN50 image encoder as the teacher model for Semantic Head distillation. The feature dimensions were set to $C_v = 2048$, $C_t = C = 1024$, with a window size of $T = 8$ and a stride of 4. Additionally, the CLIP-RN50 text encoder was borrowed to serve as the text module. The parameters of the text module were frozen throughout the entire training process to ensure the stability of text feature extraction.Both the Semantic Head and Location Head contained a five-layer MLP architecture, followed by GroupNorm and ReLU activations after the final layer. Our model was trained for 100 epochs on a platform equipped with two RTX 2080Ti GPUs. The learning rate was set to 1e-5 and a cosine annealing strategy was employed for decay.

Comparison with State-of-the-Art

Table 1 presents the performance of our method compared to the current state-of-the-art models in Refering MOT and traditional trackers on the Refer-KITTI benchmark. Most methods utilize Yolov8(Varghese and Sambath 2024) as the detector, with the Refering MOT methods including iKUN and TransRMOT, while traditional trackers are represented by ByteTrack. Due to the convenience of the iKUN architecture, we combine the iKUN model with another traditional tracker, DeepSort, where DeepSort provides the can-

Method	Detector	HOTA	DetA	AssA	DetRe	DetPr	AssRe	AssPr	MOTA	IDF1
ByteTrack(Zhang et al. 2022)	Yolov8	22.49	13.17	40.62	16.13	36.61	46.09	73.39	-7.52	23.72
TransRMOT(Wu et al. 2023)	Yolov8	38.06	29.28	50.83	40.20	47.36	55.43	81.36	9.03	46.40
Deepsort(Wojke, Bewley, and Paulus 2017)+iKUN	Yolov8	42.46	31.64	57.56	46.03	46.32	63.48	77.66	12.5	52.57
iKUN(Du et al. 2024)	Yolov8	44.56	32.05	62.48	48.53	44.76	70.52	76.66	9.69	55.40
Ours	**Yolov8**	**46.99**	**34.90**	**63.99**	**53.21**	**46.30**	**72.18**	**77.06**	**14.51**	**58.95**
iKUN*	**DeformableDETR**	**48.84**	**35.74**	**66.8**	**51.97**	**52.25**	**72.95**	**87.09**	**12.26**	**54.05**

Table 1: **The comparison with SOTA and traditional trackers.**All trackers were trained on Refer-KITTI.

Method	HOTA	AssA	MOTA	IDF1
Full Model	46.99	63.99	14.51	58.95
Without Appearance Cue	44.87	62.46	11.53	56.58
Without Motoin Cue	45.77	62.06	12.08	57.97
Without Semantic Cue	44.92	62.44	13.80	57.64

Table 2: **Effectiveness of three cues in the tracking.**

didates for the targets and trajectories of iKUN. Additionally, we combine iKUN with the more powerful object detection performance of the DeformableDETR(Zhu et al. 2020) model to demonstrate the gaps between our method and these approaches. From the experimental results, our method achieved 46.99%, 14.51%, and 58.95% on HOTA, MOTA, and IDF1, respectively. Compared to the original SOTA models, our method improved by 2.4%, 4.8%, and 3.5% on these metrics. It can be considered that our method indeed exhibits a significant enhancement over iKUN, especially with nearly a 50% increase in the MOTA metric, indicating fewer instances of missed detections, false alarms, and incorrect associations. Particularly noteworthy is that in comparison to solutions utilizing more powerful detectors, our approach is closer to iKUN in some metrics and even surpasses it in certain aspects (such as MOTA and IDF1). This exemplifies how our method can provide a more robust tracking solution.

Ablation Study

We conducted comprehensive ablation experiments to validate the effectiveness of the modules we introduced. Through three sets of distinct experiments, we verified the efficacy of different cues in the tracking process, with all results presented in Table 2, where we particularly focus on the accuracy of associations. Clearly, the performance of our method varies to different extents when any cue is removed.

When the appearance cue is lacking, there is a significant decrease in association performance, with AssA dropping by 1.53 and IDF1 decreasing by 2.37, yet still slightly outperforming the original iKUN model. This suggests the necessity of introducing appearance cues to enhance the associations between target trajectories and highlights how motion and semantic cues can also improve matching accuracy to some extent.The removal of motion cues results in the most substantial decline in AssA, indicating the importance of motion cues in the association phase. This confirms our hypothesis regarding the significance of motion cues. Additionally, the decrease in HOTA and MOTA due to the absence of semantic cues underscores the role of semantic cues in the tracking process.

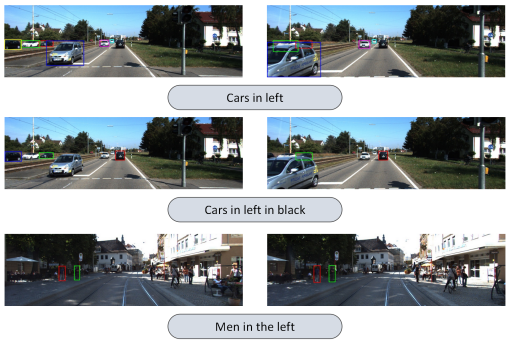

Figure 2: Qualitative results of our method on Refer-KITTI

In conclusion, the fusion of appearance, semantic, and motion cues can provide a more comprehensive understanding of object information, enabling the model to exhibit a more stable tracking performance.

Qualitative Results

We visualized some typical results (figure 2) to demonstrate the tracking effects under different text guidance. The first text focuses on the location of cars, the second text adds descriptions of car colors based on the car's position, and the third text focuses on the location of people.

Conclusion

In this work, we propose a method for multi-cue fusion in language-guided multi-object tracking, cleverly combining semantic, motion, and appearance cues for object association. This approach addresses the limitations of Refering MOT, which relies on single cues for tracking, enabling the model to leverage richer visual characteristics and contextual information to capture complex relationships during tracking. By fusing multiple cues early on, we avoid the need for complex heuristic algorithms for matching in later stages, simplifying the computational process and enhancing tracking accuracy. This method particularly demonstrates more stable performance in tracking issues within complex scenes.

References

Awais, M.; Naseer, M.; Khan, S.; Anwer, R. M.; Cholakkal, H.; Shah, M.; Yang, M.-H.; and Khan, F. S. 2023. Foundational models defining a new era in vision: A survey and outlook. *arXiv preprint arXiv:2307.13721*.

Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; and Upcroft, B. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, 3464–3468. IEEE.

Botach, A.; Zheltonozhskii, E.; and Baskin, C. 2022. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4985–4995.

Du, Y.; Lei, C.; Zhao, Z.; and Su, F. 2024. ikun: Speak to trackers without retraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19135–19144.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14084–14093.

Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; and Meng, H. 2023. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 25: 8725–8737.

Li, S.; Fischer, T.; Ke, L.; Ding, H.; Danelljan, M.; and Yu, F. 2023. Ovtrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5567–5577.

Li, Y.; Liu, X.; Liu, L.; Fan, H.; and Zhang, L. 2024a. LaMOT: Language-Guided Multi-Object Tracking. *arXiv preprint arXiv:2406.08324*.

Li, Y.; Naseer, M.; Cao, J.; Zhu, Y.; Sun, J.; Zhang, Y.; and Khan, F. S. 2024b. Multi-Granularity Language-Guided Multi-Object Tracking. *arXiv preprint arXiv:2406.04844*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Srivatsan, K.; Naseer, M.; and Nandakumar, K. 2023. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19685–19696.

Varghese, R.; and Sambath, M. 2024. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 1–6. IEEE.

Wojke, N.; Bewley, A.; and Paulus, D. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, 3645–3649. IEEE.

Wu, D.; Han, W.; Wang, T.; Dong, X.; Zhang, X.; and Shen, J. 2023. Referring multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14633–14642.

Yu, E.; Wang, T.; Li, Z.; Zhang, Y.; Zhang, X.; and Tao, W. 2023. Motrv3: Release-fetch supervision for end-to-end multi-object tracking. *arXiv preprint arXiv:2305.14298*.

Zeng, F.; Dong, B.; Zhang, Y.; Wang, T.; Zhang, X.; and Wei, Y. 2022. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, 659–675. Springer.

Zhang, H.; Wang, J.; Zhang, J.; Zhang, T.; and Zhong, B. 2023. One-stream vision-language memory network for object tracking. *IEEE Transactions on Multimedia*, 26: 1720–1730.

Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, 1–21. Springer.

Zhang, Y.; Wang, T.; and Zhang, X. 2023. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22056–22065.

Zhang, Y.; Wu, D.; Han, W.; and Dong, X. 2024. Bootstrapping Referring Multi-Object Tracking. *arXiv preprint arXiv:2406.05039*.

Zheng, Y.; Zhong, B.; Liang, Q.; Li, G.; Ji, R.; and Li, X. 2023. Towards unified token learning for vision-language tracking. *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhou, L.; Zhou, Z.; Mao, K.; and He, Z. 2023. Joint visual grounding and tracking with natural language specification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23151–23160.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.