Cross-View Referring Multi-Object Tracking

Sijia Chen, En Yu, Wenbing Tao*

Huazhong University of Science and Technology
{sijiachen, yuen, wenbingtao}@hust.edu.cn

Abstract

Referring Multi-Object Tracking (RMOT) is an important topic in the current tracking field. Its task form is to guide the tracker to track objects that match the language description. Current research mainly focuses on referring multi-object tracking under single-view, which refers to a view sequence or multiple unrelated view sequences. However, in the single-view, some appearances of objects are easily invisible, resulting in incorrect matching of objects with the language description. In this work, we propose a new task, called **C**ross-view **R**eferring **M**ulti-**O**bject **T**racking (**CRMOT**). It introduces the cross-view to obtain the appearances of objects from multiple views, avoiding the problem of the invisible appearances of objects in RMOT task. CRMOT is a more challenging task of accurately tracking the objects that match the language description and maintaining the identity consistency of objects in each cross-view. To advance CRMOT task, we construct a cross-view referring multi-object tracking benchmark based on CAMPUS and DIVOTrack datasets, named **CRTrack**. Specifically, it provides 13 different scenes and 221 language descriptions. Furthermore, we propose an end-to-end cross-view referring multi-object tracking method, named **CRTracker**. Extensive experiments on the CRTrack benchmark verify the effectiveness of our method.

Dataset, Code — https://github.com/chen-si-jia/CRMOT

Introduction

Multi-Object Tracking (MOT) is one of the most challenging tasks in computer vision. It is widely used in fields such as autonomous driving (Li et al. 2024b), video surveillance (Yi et al. 2024), and smart transportation (Bashar et al. 2022). Existing MOT methods have already demonstrated effectiveness in addressing most visual generic scenarios. However, when it comes to multimodal contexts, i.e., vision-language scenarios, traditional MOT methods face significant challenges and limitations. To solve this problem, the task of Referring Multi-Object Tracking (RMOT) was recently proposed. The form of this task is to guide the tracker to track the objects that match the language description. For example, if the input "A man in a black coat and blue trousers, carrying a blue bag and holding a book.", the network for the RMOT task will predict all target trajectories corresponding to that language description. Current research mainly focuses on the RMOT task under the single-view, which refers to a view sequence or multiple unrelated view sequences. However, in the single-view, some appearances of the objects are easily invisible, causing the network for the RMOT task to incorrectly match objects with the fine-grained language description.

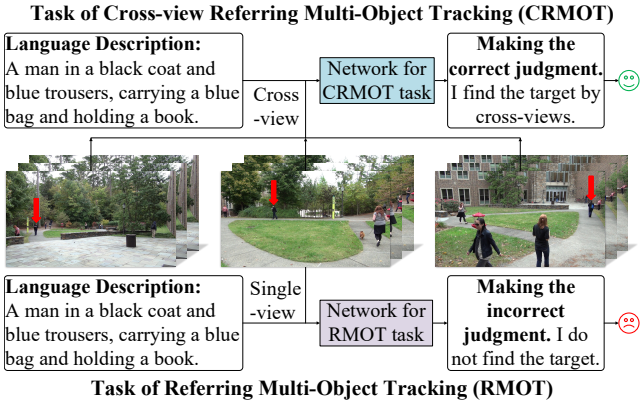

Figure 1: The difference between CRMOT and RMOT. The CRMOT task introduces the cross-view to obtain the appearances of objects from multiple views, avoiding the problem that the appearances are easily invisible in the RMOT task.

To overcome the limitation of the single-view, we propose a new task called Cross-view Referring Multi-Object Tracking (CRMOT). It introduces the cross-view, which refers to different views with large overlapping areas, to obtain the appearances of objects from multiple views, thereby avoiding the problem that the appearances of objects are easily invisible in the RMOT task. CRMOT is a more challenging task of accurately tracking the objects that match the language description and maintaining the identity (ID) consistency of the objects in each cross-view. As illustrated in Figure 1, we can observe that the network for the RMOT task makes the incorrect judgment when some appearances of the objects are invisible in the single-view of the RMOT task. In contrast, in the cross-views of the CRMOT task, the appearance of the objects can be fully captured, so that the network

*Corresponding author

for the CRMOT task can accurately track the objects that match the language description and can know which objects have the same identity (ID) in each cross-view, i.e., the network for the CRMOT task makes the correct judgment.

To advance the research on the cross-view referring multi-object tracking (CRMOT) task, we propose a benchmark, called CRTrack. Specifically, CRTrack includes 13 different scenes, 82K frames, 344 objects, and 221 language descriptions, as detailed in Table 1. These sequence scenes come from two cross-view multi-object datasets, DIVOTrack (Hao et al. 2024) and CAMPUS (Xu et al. 2016). Additionally, we propose a new annotation method based on the unchanging attributes of the objects throughout the sequences. These attributes include headwear color, headwear style, coat color and style, trousers color and style, shoes color and style, held item color, held item style, and transportation. Then, we utilize the large language model GPT-4o to generate language descriptions from the annotated attributes, followed by careful manual checking and correction to ensure the accuracy of language descriptions. Finally, we propose a set of evaluation metrics specifically designed for the CRMOT task.

Moreover, to further advance the research on the CRMOT task, we propose an end-to-end cross-view referring multi-object tracking method, called CRTracker. Specifically, CRTracker combines the accurate multi-object tracking capability of CrossMOT (Hao et al. 2024) and the powerful multi-modal capability of APTM (Yang et al. 2023). Furthermore, a prediction module is designed within the CRTracker network. The novel design idea of this prediction module is to use the frame-to-frame association results of the network as detection results, the fusion scores as confidences, and the prediction module plays the role of a tracker.

Finally, we evaluate our proposed CRTracker method and other methods on the in-domain and cross-domain test sets of the CRTrack benchmark. The evaluation results demonstrate that our method achieves state-of-the-art performance. Specifically, compared to the best-performing method among other single-view approaches, our method surpasses it by 31.45% in CVRIDF1 and 25.83% in CVRMA across all scenes in the in-domain evaluation, and by 8.74% in CVRIDF1 and 1.92% in CVRMA across all scenes in the cross-domain evaluation.

In summary, our main contributions are as follows:

1. We propose a new task, called Cross-view Referring Multi-Object Tracking (CRMOT). It is a challenging task of accurately tracking the objects that match the language description and maintaining the identity consistency of the objects in each cross-view.
2. We construct a benchmark, called CRTrack, to advance the research on the CRMOT task. This benchmark includes 13 different scenes, 82K frames, 344 objects, and 221 language descriptions.
3. We propose an end-to-end cross-view referring multi-object tracking method, called CRTracker. We evaluate CRTracker and other methods on the CRTrack benchmark both in-domain and cross-domain. The evaluation results show that CRTracker achieves state-of-the-art performance, fully demonstrating its effectiveness.

Related Work

Cross-View Multi-Object Tracking. Cross-view multi-object tracking is a specific category of multi-object tracking (Zhang et al. 2021; Zeng et al. 2022; Yu et al. 2022, 2023a,b; Chen et al. 2024; Gao, Zhang, and Wang 2024; Li et al. 2024a) that shares large overlapping areas between different views. Currently, mainstream methods (Cheng et al. 2023; Hao et al. 2024) use appearance and motion features to measure the similarity of the same pedestrians across different views and associate them. There are several commonly used cross-view multi-object tracking datasets, including DIVOTrack (Hao et al. 2024), CAMPUS (Xu et al. 2016), EPFL (Fleuret et al. 2007), WILDTRACK (Chavdarova et al. 2018), and MvMHAT (Gan et al. 2021). The DIVOTrack dataset is the latest cross-view multi-object tracking dataset with 10 scenes captured by 3 moving cameras. The CAMPUS dataset contains real scenes captured by static cameras from 3 or 4 different views. The EPFL dataset is one of the traditional cross-view tracking datasets, but its very low resolution makes it difficult to learn the appearance embeddings of objects. The WILDTRACK was shot in a square, but the pedestrian annotations are incomplete. The MvMHAT was shot on a rooftop, but all videos in the dataset use the same scene and the same person. Therefore, we choose the DIVOTrack and CAMPUS datasets to construct the benchmark.

Referring Multi-Object Tracking. Referring multi-object tracking is divided into two architectures: two-stage methods and end-to-end methods. The two-stage methods first explicitly extract object trajectories and then select object trajectories that match the language descriptions. The mainstream two-stage methods include iKUN(Du et al. 2024) and LaMOT (Li et al. 2024c). The end-to-end methods directly obtain object trajectories that match the language descriptions. The mainstream end-to-end methods include TransRMOT (Wu et al. 2023) and TempRMOT (Zhang et al. 2024).

Benchmark

To advance the research on the cross-view referring multi-object tracking (CRMOT) task, we construct a cross-view referring multi-object tracking benchmark, named CRTrack. Below, we provide details about the CRTrack benchmark.

Dataset Collection. The emphasized properties of the cross-view referring multi-object tracking dataset are two major elements: *cross-view* and *referring*. *Cross-view* refers to the overlapping area between different camera views, and *referring* refers to the language description. Therefore, based on the cross-view multi-object tracking datasets DIVOTrack (Hao et al. 2024) and CAMPUS (Xu et al. 2016), we add language descriptions to construct the cross-view referring multi-object tracking benchmark, named CRTrack. The DIVOTrack dataset contains data from 10 different real-world scenes, and it is currently the most scene-rich cross-view multi-object tracking dataset. All sequences are captured using three moving cameras and manually synchronized. The CAMPUS dataset contains 3 different scenes with frequent object occlusion problems. All sequences are captured using 3 or 4 static cameras and manually synchronized. It should be noted that we only use their training data,

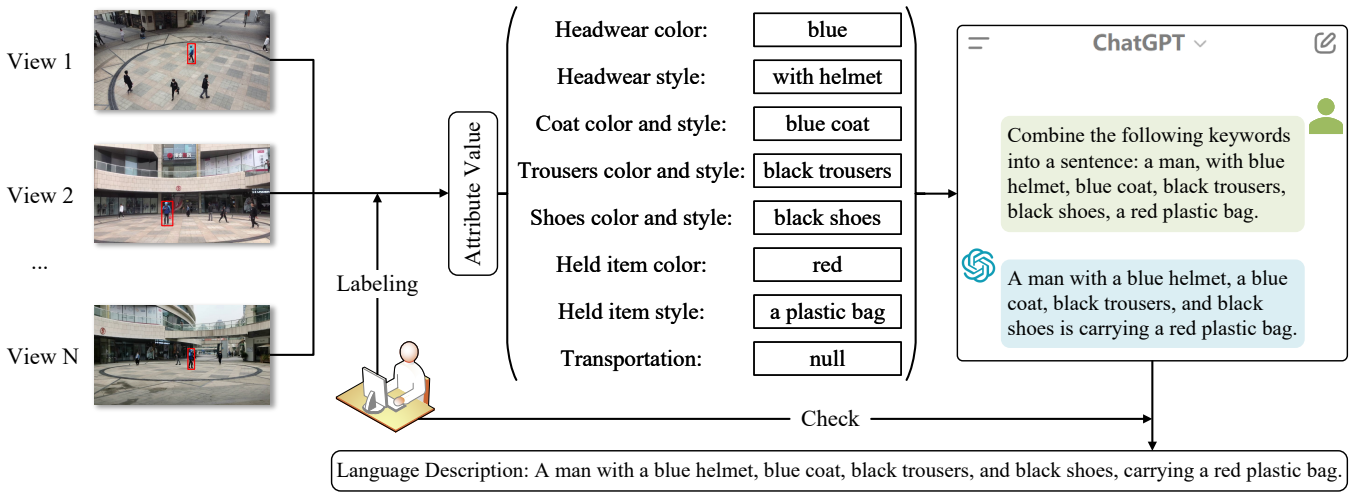

Figure 2: Language Description Annotation Pipeline.

Figure 3: Word Cloud.

and unify the image sizes and annotation formats of the DIVOTrack and CAMPUS datasets.

Dataset Annotation. We divide the content of the language description into different attributes. These attributes include headwear color, headwear style, coat color and style, trousers color and style, shoes color and style, held item color, held item style, and transportation. Detailed attributes can be found in the *supplementary materials*. Previously, some language descriptions of the RMOT task's benchmark Refer-KITTI (Wu et al. 2023) only annotate a certain fragment sequence of the object, not the whole sequence from the appearance to the disappearance of the object. This annotation method is obviously not suitable for the new task of cross-view referring multi-object tracking, because the introduction of cross-view can observe the whole sequence from the appearance to the disappearance of the object in more detail from multiple views. Therefore, we propose a new annotation method that aims to annotate objects from the perspective of their invariant attributes in the sequence, such as clothing, held items and transportation. We annotate the attributes of the objects in each scene. After obtaining the object annotation attributes, we use the large language model GPT-4o (OpenAI 2024) to produce the language descriptions based on the object annotation attributes.

Scene	Views	NFPV	Object Density	ANFLD
Floor	3	825	10.8	712
Gate1	3	1251	9.9	780
Ground	3	901	18.9	743
Moving	3	581	4.6	418
Park	3	601	7.7	599
Shop	3	1101	12.5	561
Square	3	601	9.1	457
Circle	3	1601	8.3	952
Gate2	3	801	3.6	685
Side	3	751	12.5	624
Garden1	4	2849	9.6	2742
Garden2	3	6000	5.2	3265
ParkingLot	4	6475	4.0	3419

Table 1: Dataset Statistics of the CRTrack Benchmark. NFPV represents Number of frames per view. ANFLD represents average number of frames of language descriptions.

The language descriptions generated by GPT-4o are manually checked and corrected. With the help of the large language model, the richness of the language descriptions has been greatly improved. Finally, 344 labeled objects and 221 language descriptions are obtained. The entire annotation process is shown in Figure 2.

Dataset Split. For the DIVOTrack dataset with language descriptions, we evenly selected three scenes as the in-domain test set based on the scene's object density, and the remaining seven scenes as the training set. The CAMPUS dataset with language descriptions is used as the cross-domain test set. In short, the CRTrack benchmark is divided into training set, in-domain test set and cross-domain test set. Specifically, the training set contains "Floor", "Gate1", "Ground", "Moving", "Park", "Shop" and "Square" scenes, the in-domain test set contains "Circle", "Gate2" and "Side" scenes, and the cross-domain test set contains "Garden1",

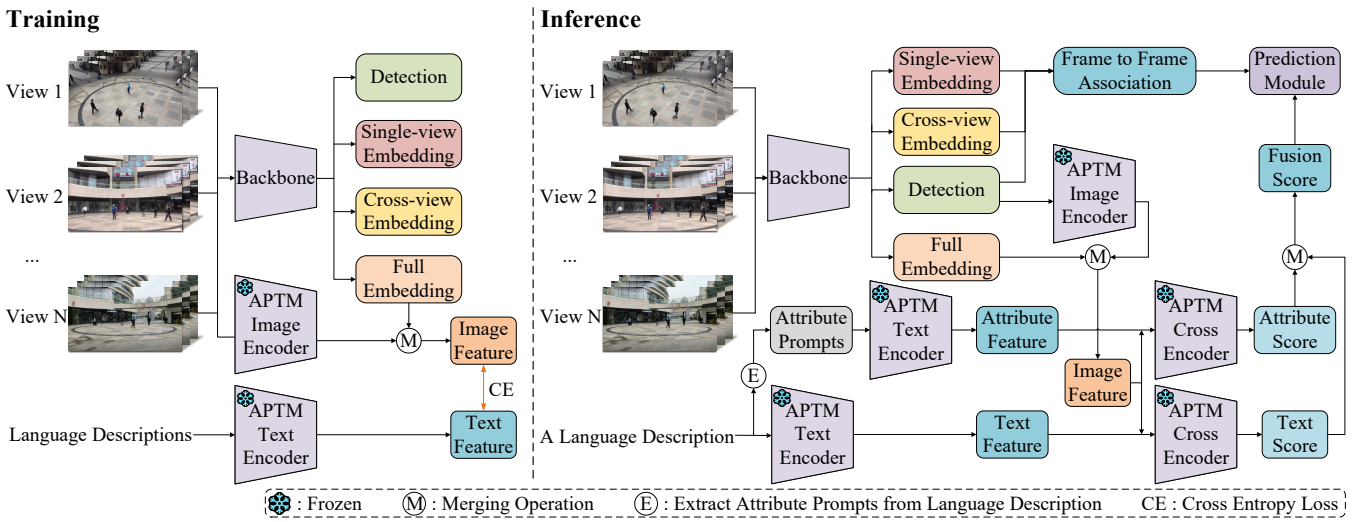

Figure 4: Pipeline of CRTracker. It includes a detection head, a single-view Re-ID head, a cross-view Re-ID head, a full Re-ID head and APTM framework. The prediction module outputs the trajectories of objects that match the language description.

"Garden2" and "ParkingLot" scenes.

Dataset Statistics. i) **Word Cloud.** Figure 3 shows the word cloud of the CRTrack benchmark we constructed. We can observe that the CRTrack benchmark contains a large number of words describing clothing, held items and transportation information. The rich variety of word clouds shows the difficulty of our benchmark. ii) **Object Density.** Object density indicates how many objects there are per frame per cross-view of a scene on average. The object density of each scene in the CRTrack benchmark is shown in Table 1. We can observe that the CRTrack benchmark has scenes with different object densities. iii) **Average Number of Frames of Language Description.** It indicates the average number of frames in which the object corresponding to each language description appears. Table 1 shows the average number of frames of the language description of each scene. The average number of frames of the language description of "ParkingLot" scene of the CRTrack benchmark reaches an astonishing 3419. The extremely long number of frames brings great challenges to the cross-view referring multi-object tracking in the temporal dimension.

Evaluation Metrics

The cross-view tracker is different from the single-view tracker. The cross-view tracker processes multiple views in each batch of synchronized video sequences. The same object should have the same identity (ID) in different views. The standard cross-view multi-object tracking evaluation metrics include the cross-view IDF1 (CVIDF1) and the cross-view matching accuracy (CVMA) (Gan et al. 2021). The definitions of CVIDF1 and CVMA are as follows:

$$CVIDF1 = \frac{2CVIDP \times CVIDR}{CVIDP + CVIDR}, \tag{1}$$

$$CVMA = 1 - \left(\frac{\sum_t m_t + fp_t + 2mme_t}{\sum_t gt_t} \right) \tag{2}$$

where CVIDP and CVIDR denote the cross-view object matching precision and recall, respectively. m_t, fp_t, mme_t, and gt_t are the numbers of misses, false positives, mismatched pairs, and the total number of objects in all views at time t, respectively.

It should be noted that cross-view referring multi-object tracking is different from cross-view multi-object tracking. When predicting non-referring but visible objects, they are considered false positives in our evaluation. When the tracking corresponding to the language description is not good, there will be a lot of false detections. This will make CVMA become a relatively large negative number, resulting in a huge impact on the evaluation metrics. We take a maximum value between CVMA value and 0 to prevent the influence of negative numbers.

We aim to comprehensively evaluate each language description, so we propose new evaluation metrics CVRIDF1 and CVRMA for the cross-view referring multi-object tracking (CRMOT) task, and their value range is 0 to 1. The definitions of the evaluation metrics CVRIDF1 and CVRMA for the CRMOT task we proposed are as follows:

$$CVRIDF1 = \frac{\sum_l CVIDF1}{n_l} \tag{3}$$

$$CVRMA = \frac{\sum_l \max(CVMA, 0)}{n_l} \tag{4}$$

where l represents a language description and n_l denotes the number of language descriptions.

Strong Baseline of CRMOT

The challenge of the CRMOT task is to simultaneously detect and track the objects that match the language description and maintain the identity consistency of the objects in each cross-view. To address the challenge of the CRMOT task, we propose an end-to-end cross-view referring multi-object tracking method, named CRTracker, as a strong baseline.

Training

APTM. APTM (Yang et al. 2023) is a framework for joint attribute prompt learning and text matching learning, including image encoder, text encoder and cross encoder. Specifically, the image encoder uses Swin Transformer (Liu et al. 2021) to output image features. The text encoder uses the first 6 layers of BERT (Devlin et al. 2018) to output text features. The cross encoder adopts the last 6 layers of BERT, fuses image features and text features, and captures semantic relationships by the cross-attention mechanism.

Pipeline of Training. The pipeline of our training framework is shown in Figure 4. The input is synchronized video sequences from multiple cross-views and language descriptions. Similar to the CrossMOT (Hao et al. 2024) algorithm, our model uses CenterNet (Zhou, Wang, and Krähenbühl 2019) as the backbone, followed by four heads, including a detection head, a single-view Re-ID head, a cross-view Re-ID head and a full Re-ID head. In addition, it also includes APTM image encoder and APTM text encoder. It is worth noting that for single-view Re-ID, the same object in different views is considered as different objects in single-view tracking; for cross-view Re-ID, the same object in different views is considered as the same object; the full Re-ID head is used for language description calculation. We use the image encoder of APTM to encode the object ground truth area in the input video sequence into the feature F_{Ai}. Then, the feature F_{Ai} is merged with the feature F_f output by the full Re-ID head to obtain the object image feature F_i. Mathematically, the merging operation can be formulated as follows:

$$F_i = F_f + \alpha F_{Ai} \tag{5}$$

where α represents the feature fusion weight of F_{Ai}.
Additionally, we use the text encoder of APTM to encode language descriptions and obtain text features. The object image features and text features are calculated using the referring loss L_r. The detection, single-view Re-ID and cross-view Re-ID are calculated using the loss L_{cmot}.

Loss Functions. Our cross-view referring multi-object tracking loss L_{crmot} is divided into two parts, cross-view multi-object tracking loss L_{cmot} and referring loss L_r.

The L_{cmot} is formulated as follows:

$$L_{cmot} = \frac{1}{2}\left(\frac{1}{e^{w_1}}L_d + \frac{1}{e^{w_2}}\left(L_s + L_c\right) + w_1 + w_2\right) \tag{6}$$

where L_d represents the detection loss, L_s represents the single-view Re-ID loss, L_c represents the cross-view Re-ID loss. w_1 and w_2 are are learnable parameters.

The L_r uses the Cross-Entropy Loss (Zhang and Sabuncu 2018), which is formulated as:

$$L_r = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{i,j}\log\left(p_{i,j}\right) \tag{7}$$

where N represents the number of objects, K represents the number of all language descriptions in the training data, $y_{i,j}$ represents the label of the j-th language description corresponding to the i-th object, $p_{i,j}$ represents the probability that the i-th object is predicted to be the j-th label value.

Thus, the final loss L_{crmot} is:

$$L_{crmot} = L_{cmot} + L_r \tag{8}$$

Algorithm 1: Prediction Module

Input: Frame-to-frame association results, i.e. input tracks of the prediction module $\mathcal{T}_{input}$; fusion scores S_f
Parameter: Fusion scores of views where the track exists $\mathcal{S}$; fusion score of the track S_f; j-th view V_j; Number of views for the track N_V; threshold of average fusion score T_{as}; threshold of single-view fusion score T_{ss}; threshold of hit score T_{hs}; hit score of the track $S^H_{\mathcal{T}_i}$; average hit score s_1; single-view hit score s_2; single-view miss score s_3
Output: Output tracks of the prediction module $\mathcal{T}_{output}$

Let $\mathcal{T}_{output} \leftarrow \emptyset$; $\mathcal{S} \leftarrow \emptyset$.
for $\mathcal{T}_i \in \mathcal{T}_{input}$ **do**
  /* summarize scores and view number of the track */
  $N_V = 0$
  for $V_j \in \{V_1, ..., V_N\}$ **do**
    if $\mathcal{T}_i$ exists in V_j **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup S_f$
      $N_V += 1$
    end if
  end for
  /* use scores to filter the track */
  if $(\sum \mathcal{S})/N_V > T_{as}$ **then**
    $S^H_{\mathcal{T}_i} += s_1$
    $\mathcal{T}_{output} \leftarrow \mathcal{T}_{output} \cup \mathcal{T}_i$
  else
    for $S_f \in \mathcal{S}$ **do**
      if $S_f > T_{ss}$ **then**
        $\lambda = \text{int}(S_f/T_{ss})$
        $S^H_{\mathcal{T}_i} += \lambda s_2$
      else
        $S^H_{\mathcal{T}_i} -= s_3$
        $S^H_{\mathcal{T}_i} = \max(S^H_{\mathcal{T}_i}, 0)$
      end if
    end for
    if $S^H_{\mathcal{T}_i} > T_{hs}$ **then**
      $\mathcal{T}_{output} \leftarrow \mathcal{T}_{output} \cup \mathcal{T}_i$
    end if
  end if
end for
return $\mathcal{T}_{output}$

Inference

Pipeline of Inference. The pipeline of our inference framework is shown in Figure 4. During the inference phase, we process language descriptions one by one. First, multiple cross-view video sequences are input into the network, and the detection head outputs the object bounding boxes. Each bounding box is matched with the corresponding single-view Re-ID features, cross-view Re-ID features, and full Re-ID features. In the frame-to-frame association step, we use the MvMHAT (Gan et al. 2021) to associate between frames and multiple cross-views. Next, we use the APTM image encoder to encode the object bounding box areas and obtain the encoded features F_{Ai}. The encoded features F_{Ai} is fused with the full Re-ID features F_f according to Formula (5) to generate the object image features F_i. Subsequently,

In-domain Evaluation										
Method	Published	Epochs	All scenes		Circle		Gate2		Side	
			CR1↑	CRA↑	CR1↑	CRA↑	CR1↑	CRA↑	CR1↑	CRA↑
TransRMOT (Wu et al. 2023)	CVPR2023	20	23.30	8.03	18.85	6.94	68.03	28.51	14.33	2.65
TransRMOT (Wu et al. 2023)	CVPR2023	100*	17.72	5.17	16.74	5.52	33.03	14.87	13.92	1.48
TempRMOT (Zhang et al. 2024)	arXiv2024	20	22.18	8.62	17.53	5.66	62.08	36.00	15.09	3.43
TempRMOT (Zhang et al. 2024)	arXiv2024	60*	23.43	10.14	20.26	8.49	63.86	45.18	14.17	0.65
CRTracker(Ours)	-	20	**54.88**	**35.97**	**58.38**	**42.44**	**91.60**	**73.40**	**37.97**	**14.87**
Cross-domain Evaluation										
Method	Published	Epochs	All scenes		Garden1		Garden2		ParkingLot	
			CR1↑	CRA↑	CR1↑	CRA↑	CR1↑	CRA↑	CR1↑	CRA↑
TransRMOT (Wu et al. 2023)	CVPR2023	20	3.66	0.20	2.85	0.01	4.23	0.55	3.87	0
TransRMOT (Wu et al. 2023)	CVPR2023	100*	2.15	0	2.22	0	2.23	0	1.97	0
TempRMOT (Zhang et al. 2024)	arXiv2024	20	3.78	0.39	3.86	0.29	2.91	0.65	4.68	0.19
TempRMOT (Zhang et al. 2024)	arXiv2024	60*	2.68	0.40	2.20	0	2.17	0.75	3.77	0.42
CRTracker(Ours)	-	20	**12.52**	**2.32**	**14.96**	**2.77**	**11.87**	**2.80**	**10.66**	**1.30**

Table 2: Quantitative results on the in-domain and cross-domain test sets of the CRTrack benchmark with CVRIDF1 ("CR1") and CVRMA ("CRA"). * Indicates that the epoch is the epoch of training in the author's paper. ↑ indicates that higher score is better. The best results are marked in **bold**.

the APTM text encoder is used to encode the input language description to obtain the text feature F_{At}. Attribute prompts are extracted from the language description and input into the APTM text encoder to obtain the attribute features F_{Aa}. Next, the APTM cross encoder is used to process the attribute features F_{Aa} and the image features F_i, and the attribute scores S_a is obtained through the head; the APTM cross encoder is used to process the text feature F_{At} and the image features F_i, and the text scores S_t is obtained through the head. Then, the text scores S_t is merged with the attribute scores S_a to obtain the fusion scores S_f. Mathematically, the merging operation can be formulated as follows:

$$S_f = S_t + \beta e^{S_a} \tag{9}$$

where β represents the score fusion weight of e^{S_a}.

Finally, the frame-to-frame association results and the fusion scores are input into the prediction module to generate the trajectories of objects that match the language description.

Prediction Module. The design idea of the prediction module is to regard the frame-to-frame association results as the detection results, the fusion scores S_f as the confidences, and the prediction module plays a tracking role. The algorithm of the prediction module is shown in Algorithm 1.

Experiments

Settings

For evaluation, we conduct experiments on the CRTrack benchmark we constructed and follow its evaluation metrics. Our models are trained for 20 epochs and tested on a single NVIDIA RTX 3090 GPU. The feature dimensions of single-view embedding, cross-view embedding, and full embedding are all set to 512. During the training phase, we use the Adam optimizer (Kingma and Ba 2014), the initial learning rate is set to 1×10^{-4}, the batchsize to 12, and the feature fusion weight α in Formula (5) to 0.01. During the inference phase, we set the score fusion weight β in Formula (9) is set to 0.1, threshold of average fusion score T_{as} to 0.5, threshold of single-view fusion score T_{ss} to 0.75, threshold of hit score T_{hs} to 30, average hit score s_1 to 3, single-view hit score s_2 to 3, and single-view miss score s_3 to 1.

Quantitative Results

On the CRTrack benchmark, we compared our CRTracker with other methods. Since previous referring multi-object tracking methods are designed for single-view, they cannot be used for the cross-view referring multi-object tracking task. Thus, we combine previous referring multi-object tracking methods with the MvMHAT (Gan et al. 2021) cross-view association algorithm to enable them to be used in the cross-view referring multi-object tracking task. In addition, since our method is end-to-end, for fair comparison, we choose two end-to-end referring multi-object tracking methods, including TransRMOT (Wu et al. 2023) and TempRMOT (Zhang et al. 2024). For in-domain evaluation, all methods are trained using the benchmark training set and tested on the benchmark in-domain test set. For cross-domain evaluation, all methods are trained using the benchmark training set and tested on the benchmark cross-domain test set. It is worth noting that our CRTracker and other methods use the same model and parameter settings for cross-domain evaluation as for in-domain evaluation.

In-domain Evaluation. As shown in Table 2, our CRTracker achieves 54.88% CVRIDF1 and 35.97% CVRMA on all scenes of the in-domain test set. In particular, it achieves 91.60% CVRIDF1 and 73.40% CVRMA on the "Gate2" scene. Notably, our CRTracker far outperforms all

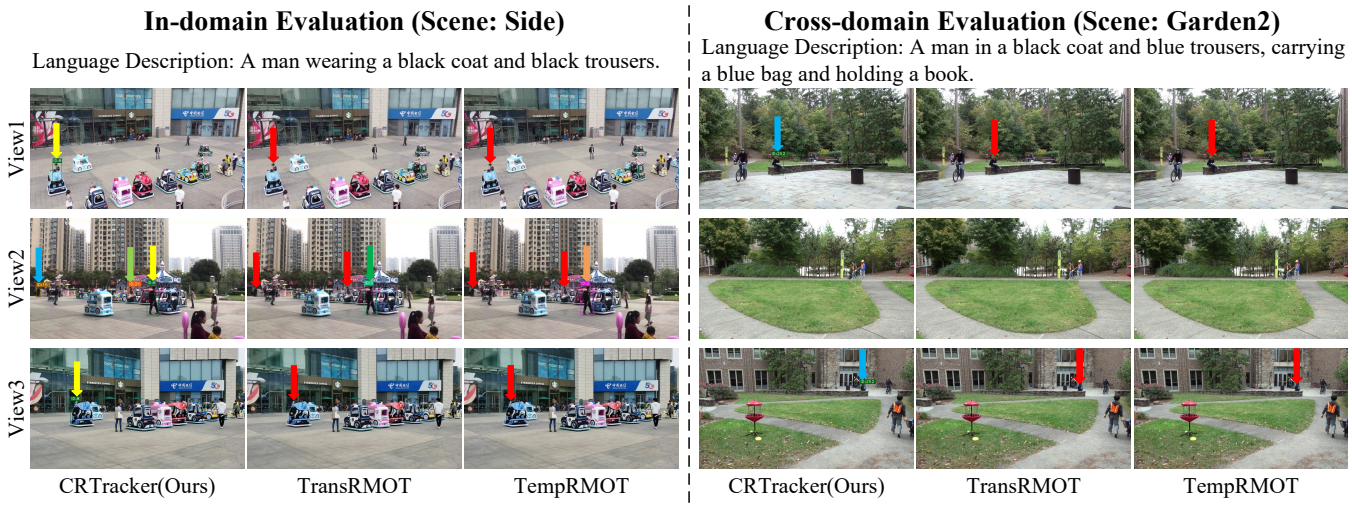

Figure 5: Qualitative results of our proposed CRTracker method and other methods, including TransRMOT and TempRMOT, on the CRTrack benchmark's in-domain and cross-domain evaluations. The rows and columns represent the camera views and different methods, respectively. Red arrows indicate targets that are not correctly detected or matched. Other colored arrows represent correctly detected targets, with arrows of the same color indicating the same target.

other methods in the in-domain evaluation. The results indicate that CRTracker can tackle in-domain scenes well.

Cross-domain Evaluation. As illustrated in Table 2, all methods suffer vital performance degradation, which is expected due to the high difficulty of the cross-domain test set of the benchmark. The cross-domain test set and the training set differ in terms of the number of cross views, scenes, pedestrians, camera angles, and lighting. In addition, the cross-domain test set contains many language descriptions that do not appear in the training set, and the average number of frames of language descriptions is very long. Despite this, our CRTracker still surpasses other methods, with achieving 12.52% CVRIDF1 and 2.32% CVRMA on all scenes of the cross-domain test set. The results show that CRTracker has a good generalization ability for unseen domains.

Qualitative Results

To further demonstrate the superiority of our CRTracker, we visualize some results of our proposed CRTracker method and other methods trained for 20 epochs in in-domain and cross-domain evaluations. As shown in Figure 5, CRTracker is able to accurately detect and track the objects that match the language description in a variety of challenging scenes and keep the same object with the same identity in each cross-view. In the "Garden2" scene example, CRTracker can accurately detect and track the target and keep the target with the same identity in each cross-view even with the untrained language description, which fully demonstrates the generalization capability of our method. Many qualitative results can be found in the *supplementary materials*.

Ablation Study

To study the role of each part of our method CRTracker, we conduct ablation experiments on the CRTrack benchmark.

Prediction Module	CVRIDF1↑	CVRMA↑
✘	47.54	28.68
✔	**54.88**	**35.97**

Table 3: Results of CRTracker without and with the prediction module. The best results are marked in **bold**.

All experiments follow in-domain evaluation, that is, training on the training set and testing on the in-domain test set.

Analysis of Prediction Module. As shown in Table 3, we can observe that the CRTracker with prediction module is 7.34% higher in CVRIDF1 and 7.29% higher in CVMA than the CRTracker without prediction module. This phenomenon shows that the prediction module fully fuses the trajectory and language description scores from each cross-view to maximize the matching of trajectory to description.

Conclusion

In this work, we propose a novel task, named Cross-view Referring Multi-Object Tracking (CRMOT). It is a challenging task of accurately tracking the objects that match the fine-grained language description and maintaining the identity consistency of the objects in each cross-view. To advance the CRMOT task, we construct the CRTrack benchmark. Furthermore, to address the challenge of the new task, we propose CRTracker, an end-to-end cross-view referring multi-object tracking method. We validate CRTracker on the CRTrack benchmark, which achieves state-of-the-art performance and demonstrates good generalization ability.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 62176096 and 61991412.

References

Bashar, M.; Islam, S.; Hussain, K. K.; Hasan, M. B.; Rahman, A.; and Kabir, M. H. 2022. Multiple object tracking in recent times: A literature review. *arXiv preprint arXiv:2209.04796*.

Chavdarova, T.; Baqué, P.; Bouquet, S.; Maksai, A.; Jose, C.; Bagautdinov, T.; Lettry, L.; Fua, P.; Van Gool, L.; and Fleuret, F. 2018. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5030–5039.

Chen, S.; Yu, E.; Li, J.; and Tao, W. 2024. Delving into the Trajectory Long-tail Distribution for Muti-object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19341–19351.

Cheng, C.-C.; Qiu, M.-X.; Chiang, C.-K.; and Lai, S.-H. 2023. Rest: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10051–10060.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Fleuret, F.; Berclaz, J.; Lengagne, R.; and Fua, P. 2007. Multicamera people tracking with a probabilistic occupancy map. *IEEE transactions on pattern analysis and machine intelligence*, 30(2): 267–282.

Gan, Y.; Han, R.; Yin, L.; Feng, W.; and Wang, S. 2021. Self-supervised multi-view multi-human association and tracking. In *Proceedings of the 29th ACM international conference on multimedia*, 282–290.

Gao, R.; Zhang, Y.; and Wang, L. 2024. Multiple Object Tracking as ID Prediction. *arXiv preprint arXiv:2403.16848*.

Hao, S.; Liu, P.; Zhan, Y.; Jin, K.; Liu, Z.; Song, M.; Hwang, J.-N.; and Wang, G. 2024. Divotrack: A novel dataset and baseline method for cross-view multi-object tracking in diverse open scenes. *International Journal of Computer Vision*, 132(4): 1075–1090.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, S.; Ke, L.; Danelljan, M.; Piccinelli, L.; Segu, M.; Van Gool, L.; and Yu, F. 2024a. Matching Anything by Segmenting Anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18963–18973.

Li, X.; Liu, D.; Zhao, L.; Wu, Y.; Wu, X.; and Gao, J. 2024b. Fast-Poly: A Fast Polyhedral Framework For 3D Multi-Object Tracking. *arXiv preprint arXiv:2403.13443*.

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

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

OpenAI. 2024. GPT-4o: OpenAI's language model. https://openai.com/blog/hello-gpt-4o/.

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.

Xu, Y.; Liu, X.; Liu, Y.; and Zhu, S.-C. 2016. Multi-view people tracking via hierarchical trajectory composition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4256–4265.

Yang, S.; Zhou, Y.; Zheng, Z.; Wang, Y.; Zhu, L.; and Wu, Y. 2023. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4492–4501.

Yi, K.; Luo, K.; Luo, X.; Huang, J.; Wu, H.; Hu, R.; and Hao, W. 2024. Ucmctrack: Multi-object tracking with uniform camera motion compensation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6702–6710.

Yu, E.; Li, Z.; Han, S.; and Wang, H. 2022. Relationtrack: Relation-aware multiple object tracking with decoupled representation. *IEEE Transactions on Multimedia*, 25: 2686–2697.

Yu, E.; Liu, S.; Li, Z.; Yang, J.; Li, Z.; Han, S.; and Tao, W. 2023a. Generalizing multiple object tracking to unseen domains by introducing natural language representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3304–3312.

Yu, E.; Wang, T.; Li, Z.; Zhang, Y.; Zhang, X.; and Tao, W. 2023b. 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, Y.; Wang, C.; Wang, X.; Zeng, W.; and Liu, W. 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129: 3069–3087.

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

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Zhou, X.; Wang, D.; and Krähenbühl, P. 2019. Objects as points. *arXiv preprint arXiv:1904.07850*.