Enhanced Kalman with Adaptive Appearance Motion SORT for Grounded Generic Multiple Object Tracking

Duy Le Dinh Anh[1], Kim Hoang Tran[1], Quang-Thuc Nguyen[2,3], and Ngan Hoang Le[1]

[1] University of Arkansas, AR, USA
[2] Faculty of Information Technology and Software Engineering Lab, University of Science, VNU-HCM
[3] Vietnam National University, Ho Chi Minh City, Vietnam

Abstract. Despite recent progress, Multi-Object Tracking (MOT) continues to face significant challenges, particularly its dependence on prior knowledge and predefined categories, complicating the tracking of unfamiliar objects. Generic Multiple Object Tracking (GMOT) emerges as a promising solution, requiring less prior information. Nevertheless, existing GMOT methods, primarily designed as OneShot-GMOT, rely heavily on initial bounding boxes and often struggle with variations in viewpoint, lighting, occlusion, and scale. To overcome the limitations inherent in both MOT and GMOT when it comes to tracking objects with specific generic attributes, we introduce **Grounded-GMOT**, an innovative tracking paradigm that enables users to track multiple generic objects in videos through natural language descriptors. Our contributions begin with the introduction of the **G^{2}MOT dataset**, which includes a collection of videos featuring a wide variety of generic objects, each accompanied by detailed textual descriptions of their attributes. Following this, we propose a novel tracking method, **KAM-SORT**, which not only effectively integrates visual appearance with motion cues but also enhances the Kalman filter. KAM-SORT proves particularly advantageous when dealing with objects of high visual similarity from the same generic category in GMOT scenarios. Through comprehensive experiments, we demonstrate that Grounded-GMOT outperforms existing OneShot-GMOT approaches. Additionally, our extensive comparisons between various trackers highlight KAM-SORT's efficacy in GMOT, further establishing its significance in the field. Project page: `https://UARK-AICV.github.io/G2MOT`. The source code and dataset will be made publicly available.

Keywords: Generic MOT · Grounded GMOT · G2MOT · KAM-SORT

1 Introduction

Multiple Object Tracking (MOT) [4,24,48,5,50,9,34,59,52,35,55,7] plays a crucial role in dynamic scene analysis, proving essential for various critical real-world applications including surveillance, security, autonomous driving, robotics, and biology. Despite significant advancements in this field, current MOT methodologies predominantly focus on a limited set of object categories, typically emphasizing a specific area of interest, such as pedestrians [23,36,14], or objects pertinent to autonomous driving scenarios [54,6]. These approaches require a substantial amount of prior knowledge about the

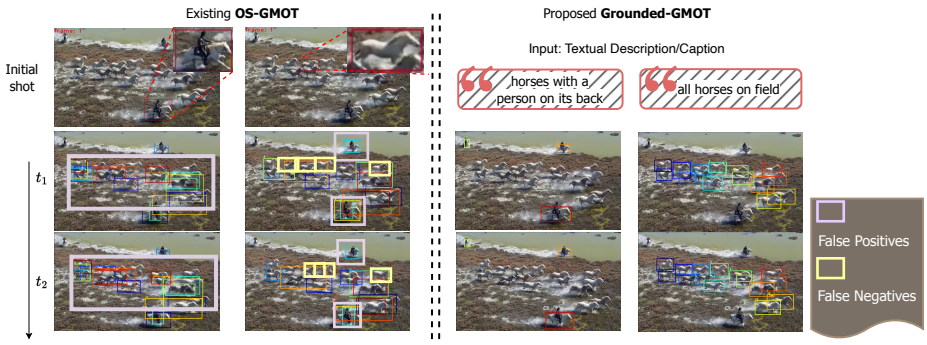


Fig. 1: Comparison between OneShot-GMOT (OS-GMOT) (left) and our Grounded-GMOT (right) in tracking multiple generic objects. The tracking system receives input (1^{st} row): OS-GMOT relies on an initial bounding box, and Grounded-GMOT utilizes textual descriptions. OS-GMOT encounters numerous challenges related to pose, illumination, occlusion, scale, texture, etc, resulting in many False Positives and False Negatives corresponding to two different initial bounding boxes. Our Grounded-GMOT adeptly detects objects based on input queries and tracks them over time at intervals t_1 and t_2 (2^{nd} and 3^{rd} rows).

objects being tracked and depend heavily on large, extensively labeled datasets. As a result, they face challenges in tracking objects across unseen or specific categories, as well as in handling objects with indistinguishable features.

Generic Multiple Object Tracking (GMOT) [32,33,2] aims to alleviate these limitations by reducing the dependency on prior information. GMOT is designed to track multiple objects of a common or similar generic type, making it suitable for a wide array of applications, ranging from annotation and video editing to monitoring animal behavior. Notwithstanding, conventional GMOT methodologies [32,33,2] are predominantly anchored in a one-shot paradigm, i.e. OneShot-GMOT (OS-GMOT), leveraging the initial bounding box of a single target object in the first frame to track all objects of the same class. The dependency on the starting bounding box poses challenges in accommodating object variations e.g., pose, illumination, occlusion, scale, texture, etc.

In recent years, significant strides have been made in achieving grounded understanding through the integration of natural language processing into computer vision [29,58,27,26]. This progress has enabled the precise alignment of language concepts with visual observations, allowing for a comprehensive understanding of both visual content and the nuance of natural language. Building upon this foundation, we aim to address the limitations of both MOT and GMOT in tracking objects with specific generic attributes. To this end, we introduce a novel tracking paradigm called **Grounded-GMOT**, which leverages the capabilities of Vision Language Models (VLMs) to guide the tracking of multiple generic objects in videos using descriptive natural language input. Figure 1 shows the comparison between OS-GMOT and our proposed Grounded-GMOT.

In this work, we first introduce **G^{2}MOT dataset**, a large-scale dataset enriched with a variety of generic object categories and their corresponding textual descriptions. G^{2}MOT dataset surpasses all currently available datasets in terms of size and diversity,

as in Table 1. We then propose **KAM-SORT** (Enhanced **K**alman Filter with Adaptive **A**ppearance **M**otion **SORT**), an innovative GMOT tracker. Our KAM-SORT tracker first enhances the Kalman Filter by integrating camera motion into the re-association of predicted bounding boxes. Subsequently, it adeptly measures appearance uniformity and dynamically adjusts the weighting between motion and appearance in the association process, ensuring robust and accurate object tracking. Our contributions are summarized as follows:

• Introducing Grounded-GMOT, a novel tracking paradigm that utilizes responsive and interactive natural language descriptions to track generic objects in videos.
• Unveiling the G^{2}MOT dataset, a novel, large-scale dataset encompassing a broad diversity of generic object categories and detailed natural language descriptions.
• Providing a comprehensive comparison between SOTA OS-GMOT and Grounded-GMOT, using a variety of popular trackers across numerous performance metrics.
• Proposing KAM-SORT, an innovative object association method that enhances the Kalman Filter and adeptly integrates both motion and appearance during the tracking.

Table 1: Comparison of **existing datasets** of SOT, MOT, GSOT, GMOT. "#" represents the quantity of the respective items. **Cat.**, **Vid.** denote Categories and Videos. **Obj.**: average number of objects per frame. **App.**: appearance similarity (%) between objects in a frame, calculated by the average cosine similarity of objects in the same frame; **Den.** density of objects in a frame, computed by the maximum number of objects at the same pixel. **Occ.**: occlusion between objects in a frame, represented by the average ratio of IoU (%) of the bounding boxes in the same frame; **Mot.**: motion speed of objects in a video, calculated by the average ratio of the IoU (%) of the bounding boxes in the same track in consecutive frames.

Datasets	Task	NLP	Statistical Information					Data Properties (mean(std))				
			#Cat.	#Vid.	#Frames	#Tracks	#Boxs	Obj.	App.	Den.	Occ.	Mot.
OTB2013 [51]	SOT	✗	10	51	29K	51	29K	–	–	–	–	–
VOT2017 [22]	SOT	✗	24	60	21K	60	21K	–	–	–	–	–
TrackingNet [39]	SOT	✗	21	31K	14M	31K	14M	–	–	–	–	–
MOT17 [37]	MOT	✗	1	14	11.2K	1.3K	0.3M	39(35)	62(10)	3.85(1.50)	14(16)	94(11)
MOT20 [14]	MOT	✗	1	8	13.41K	3.45K	1.65M	150(70)	68(8)	6.42(1.20)	15(15)	96(4)
Omni-MOT [45]	MOT	✗	1	–	14M+	250K	110M	–	–	–	–	–
DanceTrack [44]	MOT	✗	1	100	105K	990	–	9(5)	77(7)	2.67(0.99)	21(17)	90(9)
TAO [13]	MOT	✗	833	2.9K	2.6M	17.2K	333K	3(2)	69(7)	1.82(0.76)	11(14)	49(34)
SportsMOT [12]	MOT	✗	1	240	150K	3.4K	1.62M	11(3)	73(8)	2.44(0.80)	18(17)	80(16)
GOT-10 [19]	GSOT	✗	563	10K	1.5M	10K	1.5M	–	–	–	–	–
Fish [21]	GSOT	✗	1	1.6K	527.2K	8.25k	516K	–	–	–	–	–
AnimalTrack [57]	GMOT	✗	10	58	24.7K	1.92K	429K	17(9)	72(8)	3.13(1.22)	15(15)	91(11)
GMOT-40 [2]	GMOT	✗	10	40	9K	2.02K	256K	24(17)	71(9)	2.56(0.88)	11(12)	43(44)
LaSOT [16]	SOT	coarse	70	1.4K	3.52M	1.4K	3.52M	–	–	–	–	–
TNL2K [46]	SOT	coarse	–	2K	1.24M	2K	1.24M	–	–	–	–	–
Refer-KITTI [49]	MOT	coarse	2	18	6.65K	637	28.72K	5(4)	65(6)	1.78(0.74)	11(11)	73(21)
G^{2}MOT (Ours)	GMOT	fine	20	253	157.2K	5.84K	1.87M	12(5)	74(8)	2.65(0.95)	18(16)	84(14)

2 Related Work

2.1 Benchmarks

Recently, numerous benchmark datasets typically fall into two main categories: Varietal Object Tracking (VOT) [51,22,39,37,14,45,44,13,12,16,46,49] and Generic Object

Tracking (GOT) [19,21,57,2]. In VOT, the objects to be tracked typically exhibit diverse visual appearances, while in GOT, the objects share similar visual characteristics. The first category focuses on tracking a single object, encompassing Single Object Tracking (SOT) and Generic Single Object Tracking (GSOT). The second category is dedicated to tracking multiple objects, which includes Multiple Object Tracking (MOT) and Generic Multiple Object Tracking (GMOT). Table 1 shows a detailed comparison of benchmark datasets and their corresponding characteristics.

Most traditional visual tracking datasets [51,22,39,37,14,45,44,13,12,19,21,57,2] have commonly associated labels ID with individual bounding boxes. In contrast, recent tracking datasets [16,46,49] have incorporated language-assisted captions by harnessing the power of VLMs. However, existing datasets that integrate natural language processing (NLP) with textual descriptions are limited to only SOT and MOT. Our G^2MOT dataset goes beyond these limitations, supporting GMOT with rich textual descriptions and offering a significantly larger number of generic object categories and greater diversity in dataset size.

2.2 Pre-trained Vision-Language (VL) models

Recent advancements in computer vision have leveraged VL supervision, significantly enhancing model versatility and open-set recognition. A pioneering work in this domain is CLIP [41], which learns visual representations from vast amounts of image-text pairs. Since its release, CLIP has garnered attention, leading to the emergence of several VL models such as ALIGN [20], ViLD [18], RegionCLIP [62], GLIP [26,56], Grounding DINO [30], UniCL [53], X-DETR [8], OWL-ViT [38], LSeg [25], DenseCLIP [42], OpenSeg [17], and MaskCLIP [15], marking a paradigm shift across vision tasks. VL pre-training models can be categorized into three groups: (i) Image classification: Models like CLIP, ALIGN, and UniCL focus on matching images with language descriptions via bidirectional supervised contrastive learning. (ii) Object detection: This group includes ViLD, RegionCLIP, GLIPv2, X-DETR, OWL-ViT, and Grounding DINO, addressing object localization and recognition. (iii) Image segmentation: The third group involves pixel-level classification using VL models like LSeg, OpenSeg, and DenseSeg. In this work, we employ Grounding DINO as our pre-trained VL model.

2.3 Multiple Object Tracking (MOT)

Object tracking can be broadly categorized into Varietal Object Tracking, including Single Object Tracking (SOT) and MOT, and Generic Object Tracking, comprising Generic Single Object Tracking (GSOT) and GMOT. Our primary focus is on tracking multiple generic objects.

In MOT, approaches are divided based on whether detection and association are executed by a single model or separate models, known as "joint detection and tracking" and "tracking-by-detection." The first category [10,63,40,50,52,35,55] integrates detection into a single network, often with re-identification features. The second category [4,48,59,9,34] involves a two-step process: detection followed by association with previous tracklets. Tracking-by-detection has achieved SOTA results in MOT, as seen in recent studies like OC-SORT[9] and Deep-OCSORT[34].

Despite recent advancements, MOT remains tied to supervised learning and predefined categories, complicating tracking of unfamiliar objects. Unlike MOT, GMOT tracks

multiple generic objects without training data, employing a one-shot detection approach called OS-GMOT [2]. While OS-GMOT requires less prior information, it heavily relies on initial bounding boxes and struggles with variations in viewpoint, lighting, occlusion, and scale. In contrast, we introduce a novel zero-shot tracking paradigm, *Grounded-GMOT*, enabling users to track multiple generic objects in videos using natural language descriptors without prior training data or predefined categories.

Table 2: Statistical information of G^{2}MOT dataset.

Datasets	**Splits**	**Statistical Information**					**Data Properties (mean(std))**				
		#Cat.	**#Vid.**	**#Frames**	**#Tracks**	**#Boxes**	**Obj.**	**Mot.**	**Occ.**	**App.**	**Den.**
DanceTrack [44]	Train	1	40	41.8K	419	348.93K	8(5)	89(9)	20(17)	76(8)	2.62(0.99)
	Val	1	25	25.5K	273	225.15K	9(4)	91(9)	21(17)	77(6)	2.74(0.98)
AnimalTrack [57]	Train	10	32	11.5K	823	186K	16(9)	91(14)	15(15)	71(8)	3.09(1.18)
	Test	10	26	13.2K	1.1K	243K	19(7)	92(8)	15(15)	72(7)	3.17(1.26)
SportMOT [12]	Train	1	45	28.57K	639	312.58K	11(3)	80(16)	18(16)	73(8)	2.44(0.83)
	Val	1	45	26.97K	641	295.57K	11(3)	80(16)	18(17)	73(8)	2.44(0.76)
GMOT-40 [2]	Test	10	40	9.64K	1.94K	256.34K	24(17)	43(44)	11(12)	71(9)	2.56(0.88)
G^{2}MOT(Ours)	Test	20	253	157.2K	5.84K	1.87M	12(5)	84(14)	18(16)	74(8)	2.65(0.95)

3 G^{2}MOT Dataset

Ensuring a fair assessment of GMOT methods demands a dataset of consistent quality, free from annotator bias, and with a clearly defined problem setup. To offer comprehensive coverage of real-world scenarios across different research domains, our released dataset embodies two characteristics: (i) *Diversity:* integrating diverse object categories from various sources, encompassing a broad spectrum of classes and diverse properties such as motion, occlusion, appearance similarity, and density. Additionally, it employs high-level semantics like `player`, `athlete`, `referee` etc., to describe objects in complex contexts, rather than using generic terms like `person`. (ii) *Fine-Grained Annotation:* alongside capturing detailed visual attributes like color, texture, and attachments, it offers extensive textual descriptions with existing synonyms alongside captions.

3.1 Video collection

Combining datasets in object tracking offers strategic advantages. First, individual tracking datasets focus on specific challenges. Second, merging tracking datasets yields diverse challenges requiring tracking models to efficiently in varied scenarios. Therefore, by combining datasets, we can evaluate the tracking models' ability to deal with diverse scenarios e.g. object movements, density, similar appearance, and occlusion which are in line with the goal of the GMOT challenge. Finally, our ultimate objective is to *propose a new paradigm* for GMOT and create a *challenging benchmark dataset under various demanding real-world scenarios*. Our G^{2}MOT dataset, a combination four benchmarks (Table 2) not only broadens the range of object categories but also highlights the need for reliable tracking methods across different real-world contexts, e.g., *generic categories diversity, fast-moving objects, high occlusion, long gaps, camera motion, etc*.

3.2 Annotation

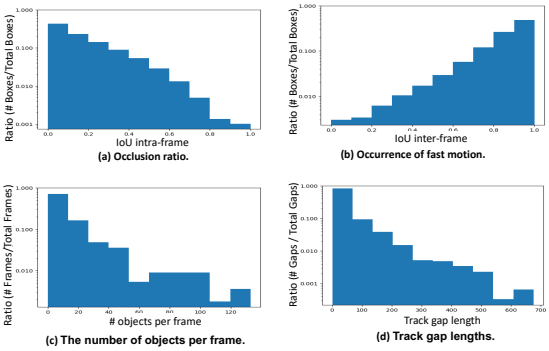
Our objective is to create a dataset with *precise descriptions* and *ambiguity-free annotations*, ensuring consistency for evaluation. Our caption generation process is *conducted manually*, emphasizing the need for careful attention to detail and accuracy. The annotation comprises two components: textual description annotation and tracking annotation.

Fig. 2: Demonstration of both "`superset`" and "`subset`" types within the same video and other fields in our annotation format which is described in Section 3.2.

Textual description annotation: The textual description annotation, as shown in Figure 2, is structured within JSON files and encompasses the following key fields: `class_name`: represents the common name of the object class; `type: superset | subset`: indicates whether the object belongs to a "superset" category, grouping "coarse category" (e.g., horse), or a "subset" category, allowing for finer categorization (e.g., horse on ground) as in Fig. 2; `caption`: manually crafted comprehensive description providing detailed information about the tracked objects; `synonyms`: Offers alternative terms or phrases for the class name; `definition`: describe the object's visual characteristics. `attributes`: encompasses a list of attributes used to distinguish objects within the "superset"; `track_path`: follows the standard MOT format and is stored separately.
Tracking annotation: The tracking annotation follows the standard MOT format [37,14] includes the following parameters [`frame_id, identity_id, box_top_left_x, box_top_left_y, box_width, box_height, 1, -1, -1, -1`].

Equipped with comprehensive annotations and diverse attributes, our G^{2}MOT dataset extends its utility beyond object tracking to support various grounding tasks e.g., Question Answering, fine-grained task understanding in real-world scenarios. Moreover, G^{2}MOT provides adaptability in tracking objects under different configurations, including captions (default), attributes, object definitions, and synonyms, effectively addressing the complexities of real-world natural language descriptions.

(a) Demonstration of word use in the caption under Word cloud in G^{2}MOT dataset.



(b) Demonstration of a wide range of statistical information of G^{2}MOT, highlighting characteristics such as occlusions (a), fast-moving objects (b), a high number of tracklets (c), and tracking gaps (d).

Fig. 3: Statistical information of our proposed G^{2}MOT.

3.3 Data statistics

Table 1 provides a comprehensive comparison with existing datasets, while Table 2 offers detailed information about our G^{2}MOT. Among GMOT datasets, ours has the highest number of categories and videos, surpassing all MOT datasets except for TAO. However, it's important to note that TAO, despite having a higher video count, lacks dense annotation and exhibits lower annotation quality, with fewer challenges such as low appearance similarity, and less occlusion. Additionally, while some datasets such as MOT20 [14] contain a high number of objects per frame (Obj.) but low appearance similarity (App.), less occlusion (Occ.), slow motion (Mot.), and DanceTrack [44] exhibit high App., substantial Occ., but slower motion (Mot.), fewer Obj., our G^{2}MOT, being a combination of multiple datasets, provides a diversity of challenges including a large Obj., high App., dense object density (Den.), substantial Occ., and fast Mot. While existing referring datasets [16,46,49] only provide captions as tracking settings and present a low range of data properties, including low scores of Obj., App., Den., Occ., and Mot., our G^{2}MOT offers fine-grained information with various textual description settings including definition, attributes, synonyms, besides captions, and contains a wide range of diversity in data properties. This is depicted in Tables 1 and 2 through the dataset's statistical information and data properties, including mean and standard deviation on each metric. Detailed computation of these metrics (Obj., App., Den., Occ., Mot.) is included in the Supp.

Figure 3(a) presents a Word Cloud that depicts the frequent usage of terms related to the caption 's *subject* and caption 's *attributes*. Not only diversity within the object category, but it also diversifies in higher semantic levels e.g., "player", "athlete", "dancer" rather than "person" only. Regarding the attribute part, the most frequently occurring descriptors encompass color, object parts, and locations. Figure 3(b) summarizes some further attributes, including (a) occlusion: measured by the IoU of interacting objects; (b)

occurrence of fast motion: determined by the IoU between object boxes in two adjacent frames; (c) number of objects per frame: calculated by number of object in a frame. (d) track gap lengths: measured by the gap length between the frames in which an object reappears and the frame of its last occurrence.

3.4 Grounded-GMOT benchmark protocols

In this evaluation process, we make use of well-established metrics as defined in [49,44]. Specifically, we employ the following metrics: Higher Order Tracking Accuracy ($HOTA$) [31], Multiple Object Tracking Accuracy ($MOTA$) [3], and $IDF1$ [43], together with Detection Accuracy ($DetA$), Association Accuracy ($AssA$). It is important to note that $HOTA = \sqrt{DetA \cdot AssA}$ effectively strikes a balance in assessing both frame-level detection and temporal association performance.

4 Proposed KAM-SORT

MOT is primarily designed to track objects with diverse appearances, such as individuals wearing various outfits or having distinct facial features and hairstyles. In contrast, GMOT is tailored for tracking objects of a generic type that share a high degree of visual similarity. This task becomes particularly challenging when attempting to associate objects across consecutive frames, especially in scenarios where objects densely congregate, as seen in groups like schools of fish, ant colonies, or swarms of bees. Consequently, our approach advocates for the utilization of both visual representations and motion cues to effectively track these generic objects. Our proposed KAM-SORT consists of two major contributions: (i) propose a tracking mechanism that can dynamically balance visual appearance and motion cues (ii) propose **Kalman++**, an improvement to the Kalman filter to re-associate unmatched detections and unmatched tracks.

Data:
$\mathcal{D}, \mathcal{T}$: set of detection boxes at current frame and tracks at the previous frame.
α: param. of uncertainty revise factor.
Model:
C: score matrix defined in Equation 5; M: bipartite matching function; K_p, K_u: Kalman Filter predict and update; BC, IoU: function compute box center and IoU.
Output:
$\mathcal{T}'$ set of new tracks.
$\hat{x}, P = K_p(\mathcal{T})$; // Get estimated location and error covariance.
$S = C(\hat{x}, \mathcal{D})$; // Compute matching score between estimation and detection.
$DT_m, D_r, T_r = M(S)$; // 1^{st}-round association produce matched pairs DT_m, unmatched detections D_r, and unmatched tracks T_r.
$S_{IoU} = IoU(D_r, T_r)$; // 2^{nd}-round associate unmatched ones.
$DT_r = M(S_{IoU})$; // Rematched pairs from remaining detections and tracks.
for $(i_d, i_t) \in DT_r$ **do**
  // i_d: detection index, i_t: track index.
  $c^{min} = \hat{x}_{i_t}[:2] - \alpha\sqrt{P[:2]}$ and $c^{max} = \hat{x}_{i_t}[:2] + \alpha\sqrt{P[:2]}$; and $c = BC(\mathcal{D}_{i_d})$;
  if $c > c^{min} \& c < c^{max}$ **then**
    $DT_m = DT_m \cup (i_d, i_t)$;
  end
end
$\mathcal{T}' = K_u(DT_m)$ // Update matched tracks

Algorithm 1: Kalman++ algorithm

The problem's setting is as follows: Consider a set of N existing tracks $\mathcal{T}$ and a set of M new detections in the current time step $\mathcal{D}$. The standard similarity between the track $\mathcal{T}$ and box embeddings $\mathcal{D}$ is defined by cosine distance and is represented as $C_a \in \mathbb{R}^{M \times N}$. In a typical tracking approach that combines visual appearance and motion cues, the cost matrix C is:

$$C(\mathcal{T}, \mathcal{D}) = C_m(\mathcal{T}, \mathcal{D}) + \gamma C_a(\mathcal{T}, \mathcal{D}). \quad (1)$$

Here, C_m represents the motion cost, which is measured using the IoU cost matrix. Leveraging OC-SORT, a technique that calculates a virtual trajectory over occlusion periods to correct error accumulation in filter parameters during occlusions, the motion cost is defined as:

$$C_m(\mathcal{T}, \mathcal{D}) = IoU(\mathcal{T}, \mathcal{D}) + \lambda C_v(\tilde{T}, \mathcal{T}). \tag{2}$$

Therefore, the resulting cost matrix that integrates both visual appearance and motion information is as follows:

$$C(\mathcal{T}, \mathcal{D}) = IoU(\mathcal{T}, \mathcal{D}) + \lambda C_v(\tilde{T}, \mathcal{D}) + \gamma C_a(\mathcal{T}, \mathcal{D}). \tag{3}$$

, where $\tilde{T}$ contains the trajectory of observations of all existing tracks. C_v represents the consistency between the directions of i) linking two observations (i.e., $\mathcal{T}^{t-2}$, $\mathcal{T}^{t-1}$: a set of tracks at time $(t-2)$ $(t-1)$), denote as, $(\mathcal{T}^{t-2} \rightarrow \mathcal{T}^{t-1})_{[u,v]}$ and ii) linking tracks' historical observations $\mathcal{T}^{t-1}$ and new observations $\mathcal{D}$ at frame t, $(\mathcal{T}^{t-1} \rightarrow \mathcal{D})_{[u,v]}$. As a result, C_v is computed on 2D coordinates $[u, v]$ of the object center:

$$C_v = \texttt{arctan}\left((\mathcal{T}^{t-2} \rightarrow \mathcal{T}^{t-1})_{[u,v]}, (\mathcal{T}^{t-1} \rightarrow \mathcal{D})_{[u,v]}\right). \tag{4}$$

To strike a balance between visual appearance and motion cues, we incorporate adaptive appearance cost W_a and adaptive motion cost W_m into Equation 3, resulting in:

$$C(\mathcal{T}, \mathcal{D}) = W_m IoU(\mathcal{T}, \mathcal{D}) + \lambda C_v(\tilde{T}, \mathcal{D}) + W_a C_a(\mathcal{T}, \mathcal{D}). \tag{5}$$

To effectively handle the high similarity between objects of the same generic type in GMOT, we propose the following hypothesis: when the visual appearances of all detections are very similar, the tracker should prioritize motion over appearance. The homogeneity of visual appearances across all detections can be quantified as follows:

$$\mu = \frac{1}{M}\sum_{i=1}^{M} f_i \text{ and } \mu_{det} = \frac{1}{M}\sum_{i=1}^{M} \texttt{cos}(f_i, \mu). \tag{6}$$

Here, we consider a threshold θ to determine the similarity between two vectors; if the angle between them is smaller than θ, the vectors are considered more similar. It's noteworthy that when $\mu_{det} > \texttt{cos}(\theta)$, the visual appearance is less reliable for tracking, implying that C_a should be less than 1. Conversely, $C_a > 1$ when $\mu_{det} < \texttt{cos}(\theta)$. Therefore, the weight C_a can be calculated as:

$$W_a = \frac{(1 - \mu_{det})}{1 - \texttt{cos}(\theta)}. \tag{7}$$

We initialize C_m as 1, indicating that both motion and appearance are equally important. As the weight on appearance reduces, we propose redistributing the remaining weight to motion. Thus, the adaptive motion weight C_m is:

$$W_m = 1 + [1 - W_a] = 2 - \frac{(1 - \mu_{det})}{1 - \texttt{cos}(\theta)}. \tag{8}$$

As a result, the final cost matrix C is computed as follows:

$$C(\mathcal{T},\mathcal{D}) = \left(2 - \frac{(1-\mu_{det})}{1-\cos(\theta)}\right) IoU(\mathcal{T},\mathcal{D}) + \lambda C_v(\tilde{T},\mathcal{D}) + \frac{(1-\mu_{det})}{1-\cos(\theta)} C_a(\mathcal{T},\mathcal{D}). \quad (9)$$

In our KAM-SORT framework, the cost matrix between existing tracks $\mathcal{T}$ and detections $\mathcal{D}$ is computed using our proposed Kalman++ algorithm, outlined in Algorithm 1. Specifically, we introduce an uncertainty revision parameter (α) to re-associate unmatched detections and tracks. From our observations, we have noticed a significant variation in box centers when dealing with fast motion and object deformation. This variation introduces unwanted noise in linear estimators like the standard Kalman Filter, leading to mismatches between detections and tracks. As a result, we propose to employ IoU scores to associate detections with previously unmatched tracks. Our Kalman++ algorithm strategically adjusts the probabilistic bounding box by considering the variance in predictions. This adjustment allows for the expansion or contraction of the predicted bounding box based on the variance of the prediction, providing flexible thresholds (c_{min}, c_{max}) that adapt to the level of uncertainty in the prediction.

5 Experimental Results

5.1 Implementation Details

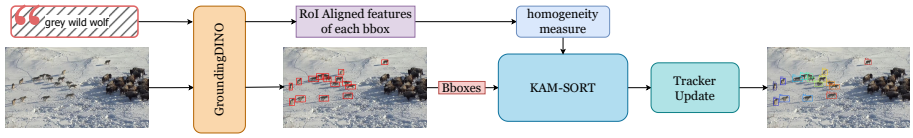


Fig. 4: Overview of the proposed Grounded-GMOT pipeline. The system detects objects using Grounding DINO with natural language descriptors (e.g., "grey wild wolf"), measure homogeneity via RoI-aligned features, and uses KAM-SORT for object association, enabling zero-shot multi-object tracking without prior training.

In the context of GMOT, the state-of-the-art (SOTA) approach is known as OS-GMOT [2]. To ensure fair comparisons, Grounded-GMOT is evaluated against OS-GMOT. We implement OS-GMOT following the configuration outlined in SOTA [2]. For Grounded-GMOT, we utilize GroundingDINO [30] alongside `captions` from the proposed G^{2}MOT dataset to generate detected bounding boxes. As shown in Fig. 4, the system detects objects using natural language descriptors (e.g., "grey wild wolf"), calculate homogeneity across objects via RoI-aligned features, and passes them to KAM-SORT for object association across frames. OS-GMOT operates on a one-shot basis, whereas our Grounded-GMOT employs a zero-shot tracking mechanism, with no training required in either OS-GMOT or Grounded-GMOT.

To evaluate the efficacy of our KAM-SORT, we compared it with several established trackers, including SORT [4], DeepSORT [48], BYTETrack [59], OC-SORT [9], Deep OCSORT [34], MOTRv2 [61]. It is worth noting that while KAM-SORT is SORT-based, MOTRv2 utilizes a transformer-based architecture. To implement KAM-SORT, we configured the parameter with uncertainty revision $\alpha = 1$ and similarity threshold $\theta = 80°$ in our experiments.

5.2 Performance Comparison

Table 3: *Tracking performance* comparison of multiple trackers under various settings of MOT with YOLOv8 [1], OS-GMOT (averaged over five runs), and our proposed *Grounded-GMOT* on the G^{2}MOT dataset. The best score is in **bold**

Trackers	Settings		HOTA↑	MOTA↑	IDF1↑	DetA↑	AssA↑
SORT [4]	YOLOv8	Fully-train	5.48	-145.61	0.80	5.78	6.47
	OS	Five runs of OS	24.77	7.09	24.90	30.22	20.70
	Grounded-GMOT	**Zero-shot**	**40.73**	**46.57**	**44.52**	**45.13**	**37.26**
DeepSORT [48]	YOLOv8	Fully-train	5.21	-156.2	0.74	5.88	5.82
	OS	Five runs of OS	22.59	-0.20	21.66	29.3	17.89
	Grounded-GMOT	Zero-shot	**36.01**	**43.30**	**37.54**	**43.94**	**29.96**
ByteTrack [59]	YOLOv8	Fully-train	6.02	-140.81	0.84	5.80	7.53
	OS-GMOT	Five runs of OS	25.16	8.02	26.46	29.38	21.94
	Grounded-GMOT	**Zero-shot**	**39.89**	**45.83**	**45.65**	**43.35**	**37.12**
OC-SORT [9]	YOLOv8	Fully-train	5.48	-127.3	0.76	5.53	6.78
	OS-GMOT	Five runs of OS	25.17	12.62	25.96	29.66	21.67
	Grounded-GMOT	**Zero-shot**	**41.84**	**46.32**	**45.92**	**44.49**	**39.92**
Deep OCSORT [34]	YOLOv8	Fully-train	5.72	-145.6	0.81	5.80	6.94
	OS-GMOT	Five runs of OS	25.65	7.06	25.92	30.47	21.92
	Grounded-GMOT	**Zero-shot**	**40.53**	**46.12**	**43.08**	**46.01**	**36.27**
MOTRv2 [61]	YOLOv8	Fully-train	3.06	0.48	0.85	0.45	20.71
	OS-GMOT	Five runs of OS	28.69	14.18	29.43	26.32	34.88
	Grounded-GMOT	**Zero-shot**	**42.02**	**41.68**	**45.91**	**41.81**	**42.54**

Comparing Grounded-GMOT with OS-GMOT and full-trained trackers. We evaluate the efficacy of the Grounded-GMOT paradigm by comparing it with (i) OS-GMOT averaged over five runs and (ii) traditional MOT settings where the object detector is fully-trained, specifically YOLOv8 [1] trained on MSCOCO [28], using various trackers as detailed in Table 3. This experiment highlights Grounded-GMOT's advancements in GMOT.

Table 5: *Tracking performance* of *KAM-SORT* on G^{2}MOT with *various settings*.

Settings	HOTA↑	MOTA↑	IDF1↑	DetA↑	AssA↑
`attribute` + `class_name`	42.20	43.26	45.29	44.73	40.15
`definition`	34.04	26.45	35.83	34.00	34.49
`caption`	**43.03**	**46.60**	**47.13**	**46.05**	**40.80**

Particularly, Grounded-GMOT demonstrates superior object detection capabilities (higher MOTA and DetA scores) attributed to the robust grounding capabilities of VLM. Additionally, leveraging textual descriptions from `captions`, Grounded-GMOT reduces reliance on initial bounding boxes, a challenge in OS-GMOT. Furthermore, while YOLOv8 [1] trained on MSCOCO [28] fails to detect categories not present in the training set, Grounded-GMOT outperforms all OS-GMOT and fully-trained MOT approaches in handling the GMOT task.

Table 6: Ablation study on the *effectiveness* of *KAM-SORT* on *MOT20-testset* with MOT task. As ByteTrack, OC-SORT uses different thresholds for testset sequences with an offline interpolation procedure, we also report scores by disabling these as in ByteTrack†, OC-SORT†. As Deep OC-SORT used separated weights for YOLOX, we also report scores by retraining YOLOX on MOT20-trainset as in Deep OC-SORT†.

Trackers	HOTA↑	MOTA↑	IDF1↑
MeMOT[7]	54.1	63.7	66.1
FairMOT[60]	54.6	61.8	67.3
GSDT[47]	53.6	67.1	67.5
CSTrack[11]	54.0	66.6	68.6
ByteTrack[59]	61.3	77.8	75.2
OC-SORT[9]	62.4	75.7	76.3
Deep-OCSORT[34]	63.9	75.6	79.2
ByteTrack†[59]	60.4	74.2	74.5
OC-SORT†[9]	60.5	73.1	74.4
Deep OC-SORT†[34]	59.6	**75.3**	75.2
KAM-SORT (Ours)	**62.6**	75.2	**76.9**

Table 4: *Tracking performance* comparison between the *existing trackers* and our proposed *KAM-SORT* tracker on G^{2}MOT dataset. The best score is in **bold**.

Trackers	Settings	HOTA↑	MOTA↑	IDF1↑	DetA↑	AssA↑
SORT [4]	Grounded-GMOT	40.73	46.57	44.52	45.13	37.26
DeepSORT [48]	Grounded-GMOT	36.01	43.30	37.54	43.94	29.96
ByteTrack [59]	Grounded-GMOT	39.89	45.83	45.65	43.35	37.12
OC-SORT [9]	Grounded-GMOT	41.84	46.32	45.92	44.49	39.92
DeepOC-SORT[34]	Grounded-GMOT	40.53	46.12	43.08	46.01	36.27
MOTRv2 [61]	Partly-trained	42.02	41.68	45.91	41.81	**42.54**
KAM-SORT(Ours)	Grounded-GMOT	**43.03**	**46.60**	**47.13**	**46.05**	40.80

Compare KAM-SORT with SOTA MOT methods.

Table 4 provides a thorough comparison between our proposed KAM-SORT and SOTA existing trackers. Our method is primarily SORT-based, necessitating an evaluation against other SORT-based approaches. In this experiment, SORT-based trackers adhere to the Grounded-GMOT setting, utilizing object detections by GroundingDINO-B. However, for a comprehensive perspective, we have also included MOTRv2 [61], a cutting-edge transformer-based tracker, as a reference point in this assessment. The results demonstrate that KAM-SORT outperforms all trackers including both SORT-based and transformer-based ones across metrics except MOTRv2 on AssA score. It is noteworthy that MOTRv2 was partially trained on the G^{2}MOT dataset, specifically on the training set of DanceTrack dataset[44]. Additionally, we conduct a visual comparison between our KAM-SORT and other trackers, as depicted in Fig. 5. In this illustration, SORT encounters challenges with loss track and incorrectly re-ID, resulting in numerous new IDs being associated at Frame #90. Let's consider the object with "ID = 6" in other trackers, OC-SORT struggles with ID Re-association caused by loss track, while Deep-SORT faces issues with ID switching and incorrectly Re-ID. In contrast, our KAM-SORT accurately re-associates object ID once the object reappears.

Table 7: An ablation study conducted on the G^{2}MOT dataset to demonstrate the *impact of each proposed component* within *KAM-SORT*.

Exp.	Appearance-Motion Balance	Kaman++	Tracking Metrics				
			HOTA↑	MOTA↑	IDF1↑	DetA↑	AssA↑
#1	✗	✗	40.53	46.12	43.08	46.01	37.27
#2	✗	✓	41.90	46.35	45.27	46.02	38.71
#3	✓	✗	43.03	46.60	47.12	46.05	40.79
#4	✓	✓	43.03	46.60	47.13	46.05	40.80

Table 8: Ablation study on *hyper-parameters* on *KAM-SORT*.

Vector Similarity θ				Uncertainty Revision α			
θ	HOTA↑	MOTA↑	IDF1↑	α	HOTA↑	MOTA↑	IDF1↑
22.5°	42.963	46.586	47.013	0.5	43.026	46.601	47.123
45°	43.010	46.600	47.091	1	43.027	46.601	47.126
67.5°	43.020	46.601	47.122	2	43.026	46.602	47.131
80°	43.027	46.601	47.126	3	43.026	46.602	47.131

5.3 Ablation Study

We conducted four ablation studies as follows:

The first ablation study, presented in Table 5, evaluates KAM-SORT's tracking performance on the G^{2}MOT dataset under different annotation settings. These settings include using both `attribute` and `class_name`, using `definition`, and using `caption` as the default setting. This study highlights the utility and accuracy of our fine-grained and informative annotations, valuable not only for object tracking but also

for many other applications and further exploration. Fig.6 illustrates some visualization of tracking performance by KAM-SORT on different annotation settings. Although we do not report tracking performance using `synonym` in Table 5 due to it represents a list of synonyms, the illustration in Fig. 6 demonstrates promising results, indicating the potential for further exploration in the future.

The second ablation study, shown in Table 6, assesses KAM-SORT's performance on the MOT task by comparing it on the MOT20 dataset [14]. To ensure fairness, we disable certain ad-hoc settings, such as employing varying thresholds for individual sequences, an offline interpolation procedure, and internal weights for object detection. YOLOX object detector is used for all trackers to demonstrate the effectiveness of KAM-SORT.

The third ablation study, presented in Table 7, evaluates two novel components of KAM-SORT: (a) Appearance-Motion Balance, which balances visual appearance and motion cues (Eq. 5), and (b) Kalman++ (Algorithm 1), an alternative algorithm replacing traditional Kalman. The ablation study (`#1` v.s. `#2`) and (`#3`v.s. `#4`) shows the importance impact of Kaman++ whereas (`#1` v.s. `#3`) and (`#2` v.s. `#4`) shows the importance impact Appearance-Motion Balance.

The fourth ablation study, shown in Table 8, report KAM-SORT's tracking performance on various hyper-parameters, i.e. vector similarity θ, defined in Eq. 7 and uncertainly revision α, defined in Algorithm 1.

Fig. 5: Tracking comparison between our tracker KAM-SORT (1^{st} column) with SORT (2^{nd} column), OC-SORT (3^{rd} column) and DeepOCSORT (4^{th} column) on video dancetrack0010 [44]. When handling objects disappear and reappear, SORT encounters challenges in maintaining tracklets, OC-SORT tends to lose track, potentially leading to incorrect Re-ID, and DeepOCSORT faces difficulties Re-ID objects once they reappear. In contrast, our KAM-SORT accurently re-associates object ID once the object reappears.

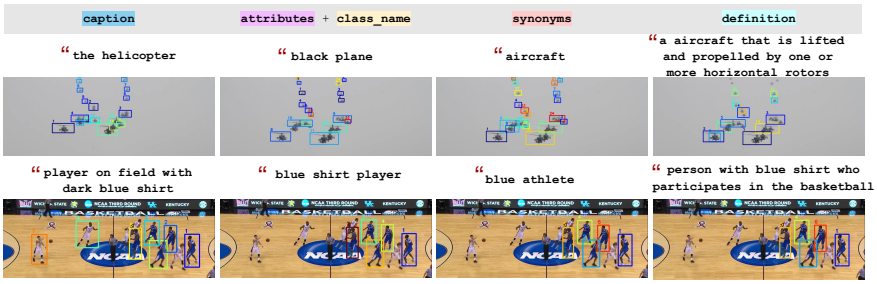


Fig. 6: KAM-SORT's tracking performance on the G^{2}MOT dataset under different textual description settings. From left to right: detected boxes and tracking IDs when using `caption` (default setting), `attribution` + `class_name`, `synonyms` (randomly select one synonym from a list of synonyms), `definition`.

6 Conclusion and Discussion

In this paper, we introduced a novel generic multi-object tracking (GMOT) framework called **Grounded-GMOT**, which leverages natural language descriptions for tracking multiple generic objects in videos. Alongside this framework, we unveiled the G^{2}MOT dataset, providing diverse object categories, substantial data size, and challenging properties including a large number of objects, fast motion, large occlusion, high appearance similarity, and high object density. The **G^{2}MOT** dataset is annotated with fine-grained language descriptions including `synonyms`, `descriptions`, `attributes`, `definitions`, and `captions`. Additionally, we presented **KAM-SORT**, an innovative object association method that incorporates Kaman++, an enhancement of the Kalman Filter and effectively balancing between motion and appearance cues. Our extensive experiments demonstrated the remarkable efficacy of the Grounded-GMOT framework in GMOT task, significantly outperforming existing SOTA OS-GMOT methods. Furthermore, our experiments and ablation studies highlighted KAM-SORT's superior performance compared to all SOTA trackers in both GMOT and MOT tasks.

Discussion: In our Grounded-GMOT framework, we utilize Grounding DINO as our preferred VLM for detecting object bounding boxes, using textual description `captions` as the query input as default. However, it is important to recognize the rich diversity of VLMs available in the field, which opens up exciting avenues for deeper exploration. Researchers and users have the opportunity to explore many other alternative VLMs specifically designed for object detection, including noteworthy options like ViLD, RegionCLIP, GLIP, X-DETR, and OWL-ViT (Section 2.2). Moreover, in our ablation study, we implemented Grounded-GMOT using various textual description settings including `definitions`, `synonym`, `attribute + class_name`, in addition to the default `caption`. This showcases the informative annotation of our proposed fine-grained G^{2}MOT dataset, which holds potential for various research and future exploration endeavors. Exploring these additional aspects of G^{2}MOT could lead to enhanced object tracking capabilities and advancements in fields such as surveillance, robotics, and animal welfare.

References

1. Yolov8: https://github.com/ultralytics/ultralytics
2. Bai, H., Cheng, W., Chu, P., Liu, J., Zhang, K., Ling, H.: Gmot-40: A benchmark for generic multiple object tracking. In: CVPR. pp. 6719–6728 (2021)
3. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The clear mot metrics. EURASIP Journal on Image and Video Processing **2008**, 1–10 (2008). https://doi.org/10.1155/2008/246309, http://dx.doi.org/10.1155/2008/246309
4. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: ICIP. pp. 3464–3468. IEEE (2016)
5. Brasó, G., Leal-Taixé, L.: Learning a neural solver for multiple object tracking. In: CVPR. pp. 6247–6257 (2020)
6. Caesar, H., Bankiti, V., H. Lang, A., Vora, S., Liong, E.V., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR. pp. 11621–11631 (2020)
7. Cai, J., Xu, M., Li, W., Xiong, Y., Xia, W., Tu, Z., Soatto, S.: Memot: Multi-object tracking with memory. In: CVPR. pp. 8090–8100 (2022)
8. Cai, Z., Kwon, G., Ravichandran, A., Bas, E., Tu, Z., Bhotika, R., Soatto, S.: X-detr: A versatile architecture for instance-wise vision-language tasks. ECCV (2022)
9. Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K.: Observation-centric sort: Rethinking sort for robust multi-object tracking. In: CVPR. pp. 9686–9696 (2023)
10. Chan, S., Jia, Y., Zhou, X., Bai, C., Chen, S., Zhang, X.: Online multiple object tracking using joint detection and embedding network. Pattern Recognition **130**, 108793 (2022)
11. Chao, L., Zhipeng, Z., Yi, L., Xue, Z., Bing, L., Xiyong, Y., Jianxiao, Z.: Rethinking the competition between detection and reid in multi-object tracking. IEEE TIP (2022)
12. Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L.: Sportsmot: A large multi-object tracking dataset in multiple sports scenes. arXiv preprint arXiv:2304.05170 (2023)
13. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: Tao: A large-scale benchmark for tracking any object. In: ECCV. pp. 436–454. Springer (2020)
14. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020)
15. Ding, Z., Wang, J., Tu, Z.: Open-vocabulary panoptic segmentation with maskclip. arXiv preprint arXiv:2208.08984 (2022)
16. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: CVPR. pp. 5374–5383 (2019)
17. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Open-vocabulary image segmentation. ECCV (2022)
18. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. ICLR (2022)
19. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE TPAMI **43**(5), 1562–1577 (2019)
20. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICLR. pp. 4904–4916. PMLR (2021)
21. Kay, J., Kulits, P., Stathatos, S., Deng, S., Young, E., Beery, S., Van Horn, G., Perona, P.: The caltech fish counting dataset: A benchmark for multiple-object tracking and counting. In: ECCV. pp. 290–311. Springer (2022)
22. Kristan, M., Matas, J., Leonardis, A., Vojíř, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., Čehovin, L.: A novel performance evaluation methodology for single-target trackers. IEEE TPAMI **38**(11), 2137–2155 (2016)

23. Leal-Taixé, L., Milan, A., et al.: MOTChallenge 2015: Towards a benchmark for multi-target tracking. arXiv:1504.01942 [cs] (2015)
24. Leal-Taixé, L., Canton-Ferrer, C., Schindler, K.: Learning by tracking: Siamese cnn for robust target association. In: CVPRW. pp. 33–40 (2016)
25. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2022)
26. Li, L.H., Zhang, P., et al.: Grounded language-image pre-training. In: CVPR. pp. 10965–10975 (2022)
27. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR. pp. 22511–22521 (2023)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context, p. 740–755. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-10602-1_48, http://dx.doi.org/10.1007/978-3-319-10602-1_48
29. Liu, F., Liu, Y., Ren, X., He, X., Sun, X.: Aligning visual regions and textual concepts for semantic-grounded image representations. Advances in Neural Information Processing Systems **32** (2019)
30. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
31. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International Journal of Computer Vision **129**(2), 548–578 (Oct 2020). https://doi.org/10.1007/s11263-020-01375-2, http://dx.doi.org/10.1007/s11263-020-01375-2
32. Luo, W., Kim, T.K.: Generic object crowd tracking by multi-task learning. In: BMVC. vol. 1, p. 3 (2013)
33. Luo, W., Kim, T.k., Stenger, B., Zhao, X., Cipolla, R.: Bi-label propagation for generic multiple object tracking. In: CVPR. pp. 1290–1297 (2014)
34. Maggiolino, G., Ahmad, A., Cao, J., Kitani, K.: Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. arXiv preprint arXiv:2302.11813 (2023)
35. Meinhardt, T., Kirillov, A., et al.: Trackformer: Multi-object tracking with transformers. In: CVPR. pp. 8844–8854 (2022)
36. Milan, A., Leal-Taixé, L., et al.: MOT16: A benchmark for multi-object tracking. arXiv:1603.00831 [cs] (2016), http://arxiv.org/abs/1603.00831
37. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
38. Minderer, M., Gritsenko, A., et al.: Simple open-vocabulary object detection with vision transformers. ECCV (2022)
39. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: ECCV. pp. 300–317 (2018)
40. Pang, J., Qiu, L., et al.: Quasi-dense similarity learning for multiple object tracking. In: CVPR. pp. 164–173 (2021)
41. Radford, A., Kim, J.W., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
42. Rao, Y., Zhao, W., et al.: Denseclip: Language-guided dense prediction with context-aware prompting. In: CVPR. pp. 18082–18091 (2022)
43. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance Measures and a Data Set for Multi-target, Multi-camera Tracking, p. 17–35. Springer International Publishing (2016). https://doi.org/10.1007/978-3-319-48881-3_2, http://dx.doi.org/10.1007/978-3-319-48881-3_2

44. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: CVPR. pp. 20993–21002 (2022)
45. Sun, S., Akhtar, N., Song, X., Song, H., Mian, A., Shah, M.: Simultaneous detection and tracking with motion modelling for multiple object tracking. In: ECCV. pp. 626–643. Springer (2020)
46. Wang, X., Shu, X., Zhang, Z., Jiang, B., Wang, Y., Tian, Y., Wu, F.: Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In: CVPR. pp. 13763–13773 (2021)
47. Wang, Y., Kitani, K., Weng, X.: Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. arXiv:2006.13164 (2020)
48. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: ICIP. pp. 3645–3649. IEEE (2017)
49. Wu, D., Han, W., Wang, T., Dong, X., Zhang, X., Shen, J.: Referring multi-object tracking. In: CVPR. pp. 14633–14642 (2023)
50. Wu, J., Cao, J., et al.: Track to detect and segment: An online multi-object tracker. In: CVPR. pp. 12352–12361 (2021)
51. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: CVPR. pp. 2411–2418 (2013)
52. Yan, B., Jiang, Y., et al.: Towards grand unification of object tracking. In: ECCV (2022)
53. Yang, J., Li, C., et al.: Unified contrastive learning in image-text-label space. In: CVPR. pp. 19163–19173 (2022)
54. Yu, F., Chen, H., Wang, X., et al.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. arXiv preprint arXiv: 1805.04687 (2018)
55. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. In: ECCV. pp. 659–675. Springer (2022)
56. Zhang, H., Zhang, P., Hu, X., Chen, Y.C., Li, L.H., Dai, X., Wang, L., Yuan, L., Hwang, J.N., Gao, J.: Glipv2: Unifying localization and vision-language understanding. NIPS (2022)
57. Zhang, L., Gao, J., Xiao, Z., Fan, H.: Animaltrack: A benchmark for multi-animal tracking in the wild. IJCV pp. 1–18 (2022)
58. Zhang, W., Shi, H., Tang, S., Xiao, J., Yu, Q., Zhuang, Y.: Consensus graph representation learning for better grounded image captioning. In: AAAI. vol. 35, pp. 3394–3402 (2021)
59. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: ECCV (2022)
60. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. IJCV **129**, 3069–3087 (2021)
61. Zhang, Y., Wang, T., Zhang, X.: Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In: CVPR. pp. 22056–22065 (2023)
62. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., Gao, J.: Regionclip: Region-based language-image pretraining. In: CVPR. pp. 16793–16803 (2022)
63. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: ECCV. pp. 474–490 (2020)