Supplementary Materials:
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

In this supplementary, we will further explore the following aspects:

- Appendix 1 provides additional details about G^{2}MOT dataset.
- Appendix 2 presents further discussion on the proposed Kalman++ and further qualitative comparison between our KAM-SORT with SORT, OC-SORT and Deep-OCSORT.

We will release our dataset and code at:
`https://UARK-AICV.github.io/G2MOT`

1 G^{2}MOT Dataset

1.1 Data Property and Criteria Definition

GMOT is designed for tracking multiple generic objects. Therefore, our aim is to construct a large-scale dataset that not only features high similarity in appearance (i.e., generic obejcts) but also encompasses challenges such as a large number of objects per frame, high density, substantial occlusion, and fast motion.

While the the main paper presents comprehensive numerical information for each criteria, this supplementary document examines the diversity of each criterion within every dataset, as illustrated in Table I. The diversity thresholds for each criterion have been determined based on our analysis and observation.

These criteria are defined as follows:

- **Obj.**: average number of objects per frame.
  Obj. = $\frac{1}{N}\sum_{t=1}^{N} M^t$, where M^t is the number of objects of frame t^{th} and N is the number of frames.
- **Occ.(%)**: occlusion between objects in a frame, represented by the average ratio of IoU of the bounding boxes in the same frame:
  Occ = $\frac{1}{N}\sum_{t=1}^{N}\left[\frac{2}{M^t\times(M^t-1)}\sum_{i=1}^{M^t-1}\sum_{j=i+1}^{M^t} IoU(O_i^t, O_j^t)\right]$, where O_i^t, O_j^t are two objects i^{th} and j^{th} in the frame t^{th}.
- **App.(%)**: appearance similarity between objects in a frame, calculated by the average cosine similarity of objects in the same frame.
  App = $\frac{1}{N}\sum_{t=1}^{N}\left[\frac{2}{M^t\times(M^t-1)}\sum_{i=1}^{M^t-1}\sum_{j=i+1}^{M^t} cos < F(O_i^t), F(O_j^t) >\right]$, where

Table I: **Diversity comparison** of existing MOT and GMOT datasets. The diversity of each criterion is determined by the thresholds indicated below each criterion.

Datasets	Task	NLP	Statistical Information					Data Properties				
			#Cat. >10	#Vid. >100	#Frames >100K	#Tracks >5K	#Boxs >1M	Obj. >10(>5)	App. >70(>5)	Den. >2.5(>0.5))	Occ. >15(>5)	Mot. >80(>5)
MOT17 [5]	MOT	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✓
MOT20 [4]	MOT	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓	✗
Omni-MOT [7]	MOT	✗	✗	–	✓	✓	✗	–	–	–	–	–
DanceTrack [6]	MOT	✗	✗	✓	✓	✗	–	✗	✓	✓	✓	✓
TAO [3]	MOT	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
SportsMOT [2]	MOT	✗	✗	✓	✓	✗	✗	✗	✓	✗	✓	✓
AnimalTrack [9]	GMOT	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓
GMOT-40 [1]	GMOT	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗
Refer-KITTI [8]	MOT	coarse	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
G^{2}MOT (Ours)	GMOT	fine	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

$F(O_i^t), F(O_j^t)$ are feature embeddings of two objects i^{th} and j^{th} in the frame t^{th}.

– **Den.**: density of objects in a frame, computed by the maximum number of objects at the same pixel.
Den. = $\frac{1}{N}\sum_{t=1}^{N} Max(H^t)$, where H^t is a density heatmap of the frame t^{th}, with each element $H^t[i,j]$ represents the number of objects occupying the pixel $[i,j]$. $Max(.)$ returns the maximum value.

– **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.
Mot. = $\frac{1}{\sum_{i=1}^{K}[Length(Track_i)-1]}\sum_{i=1}^{K}\sum_{t=1}^{|Length(Track_i)|-1} IoU(Track_i^t, Track_i^{t+1})$, where K is the number of tracks in the video, and $Track_i^t$ is a track corresponding to a track ID i^{th} throughout the video at frame t^{th}.

1.2 Annotation

The annotation format of JSON files in our G^{2}MOT dataset is as follows:

Listing 1.1: JSON format defined in G^{2}MOT dataset.

```
"video": [{
    "id": int,
    "video_path": str
}],
"tracking_query": [{
    "id": int,
    "video_id"; int,
    "is_eval": bool,
    "type": str, #"superset" or "subset"
    "superset_idx": int #-1 if type "superset",
    "class_name": str,
    "synonyms": [str],
    "definition": str,
    "attributes": [str],
    "track_path": str,
    "caption": str,
}]
```

A complete annotation of a video is depicted in Figure I. Additionally, Figure II provides textual descriptions for additional videos.

Table II: Statistics information of G2MOT. # denotes the quantity of the respective items.

Class Name	# Frames	# Objects	# Boxes	# superset	# subset
airplane	1181	81	23307	4	0
athlete	90642	1282	591146	45	131
balloon	3218	557	83430	4	11
ball	729	181	20399	4	8
bird	3462	307	70972	5	4
boat	1472	143	29491	4	5
car	2023	221	33893	4	7
chicken	7615	226	54710	5	19
deer	2762	222	53339	7	2
dolphin	1718	167	31112	6	0
duck	4851	277	79416	6	0
fish	569	291	21922	4	2
goose	1773	195	33120	5	0
horse	6556	281	69524	7	4
insect	659	297	14107	4	0
penguin	1844	137	30312	6	0
person	83623	939	688535	69	22
pig	1531	151	32273	5	0
player	85474	1254	615717	45	135
rabbit	1558	313	33961	5	0
stock	1128	143	34058	4	2
zebra	1378	99	22331	5	0

1.3 Statistical Information of G^{2}MOT

Table II provides the statistical information for each class object in G^{2}MOT dataset.

G^{2}MOT integrates datasets from GMOT-40, AnimalTrack, DanceTrack, and SportMOT, thereby presenting a diverse array of challenges for GMOT as outlined in Table I. The statistical details from each sub-dataset are provided in fig. III. Specifically, SportMOT emphasizes tracking across extended intervals and distinguishing between similar appearances. This task is complicated by factors such as uniform attire and the presence of small, distant objects. Notably, SportMOT exhibits the highest subset-to-superset ratio, making object localization challenging due to subtle visual discrepancies.

In addition, both SportMOT and DanceTrack datasets highlight significant movement and substantial occlusion, along with instances of high similarity in appearance. Meanwhile, GMOT-40 and AnimalTrack datasets present a multitude of small, similarly appearing objects with high density, as well as a high average object count per frame.

2 Kalman++ Discussion and Qualitative Results

2.1 Kalman++ Discussion

Object tracking is a challenging task, especially under conditions such as low frame rates, camera movement, object deformation, etc which introduce noise to estimators. Linear estimators like the Kalman Filter are particularly affected. While the Kalman Filter assesses the state uncertainty covariance, many Kalman Filter-based trackers primarily

rely on their own predictions of the next state in the future frame during the object association phase. In this process, the uncertainty covariance is overlooked and typically only involved in calculating the Kalman Gain. To address this, we introduce a second-phase matching into Kalman, creating Kalman++. This enables effective utilization of uncertainty covariance while mitigating the negative effects of the aformentioned conditions.

2.2 Qualitative Results

In addition to Figure 4 in the main paper, which primarily addresses *substantial occlusion* issues in tracking, we provide further illustrations demonstrating the effectiveness of KAM-SORT as follows:

Qualitative results on fast motion

fig. IV illustrates a qualitative comparison of *fast-moving object tracking* between our KAM-SORT and SORT, OC-SORT, and DeepOCSORT. In this scenario, the tracked objects (i.e., bees) exhibit high-speed movement, leading to blurring effects that impact both their appearance and detected bounding boxes. Consequently, issues such as lost track and incorrectly-ReID are common in SORT, OC-SORT, and DeepOCSORT.

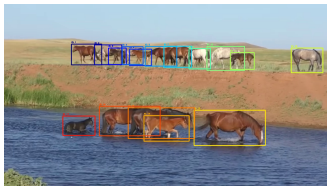
Qualitative results on high similarity in appearance

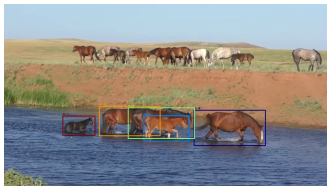
fig. V illustrates a qualitative comparison of *nearly identical objects* between our KAM-SORT and SORT, OC-SORT, and DeepOCSORT. In this scenario, the tracked objects (i.e., birds) exhibit highly similar appearances. Consequently, incorrectly re-ID is frequent in SORT, OC-SORT, and DeepOCSORT.

Qualitative results on camera motion

fig. VI illustrates a qualitative comparison of *tracking objects with camera motion* between our KAM-SORT and SORT, OC-SORT, and DeepOCSORT. In this scenario, both tracked objects (i.e., ducks) and cameras moving, which results in a big off-set between the predicted boxes and the actual bounding boxes around objects. Consequently, incorrectly re-ID is frequent in SORT, OC-SORT, and DeepOCSORT.

By flexibly balancing between appearance and motion cues, our KAM-SORT enable robust maintenance of object IDs throughout the tracking process across various conditions.






```
{
    "videos": [
        {
            "id": 1,
            "video_path": "G2MOT/AnimalTrack/horse_7"
        }
    ],
    "tracking_queries": [
        {
            "id": 1,
            "video_id": 1,
            "is_eval": true,
            "type": "superset",
            "superset_idx": -1,
            "class_name": "horse",
            "synonyms": ["pony", "equine", "steed", "charger", "mount",
            "mare", "stallion", "colt", "filly"],
            "definition": "mammal has four-legged structure, hooves, and
            distinctive characteristics such as a long mane and tail",
            "attributes": [],
            "track_path": "horse_7/query_01.txt",
            "caption": "horse"
        },
        {
            "id": 2,
            "video_id": 1,
            "is_eval": true,
            "type": "subset",
            "superset_idx": 1,
            "class_name": "horse",
            "synonyms": ["pony", "equine", "steed", "charger", "mount",
            "mare", "stallion", "colt", "filly"],
            "definition": "mammal has four-legged structure, hooves, and
            distinctive characteristics such as a long mane and tail",
            "attributes": ["on ground"],
            "track_path": "horse_7/query_02.txt",
            "caption": "horse on ground"
        },
        {
            "id": 3,
            "video_id": 1,
            "is_eval": true,
            "type": "subset",
            "superset_idx": 1,
            "class_name": "horse",
            "synonyms": ["pony", "equine", "steed", "charger", "mount",
            "mare", "stallion", "colt", "filly"],
            "definition": "mammal has four-legged structure, hooves, and
            distinctive characteristics such as a long mane and tail",
            "attributes": ["in river"],
            "track_path": "horse_7/query_03.txt",
            "caption": "horse in river"
        }
    ]
}
```


Fig. I: As an illustration of our annotation, consider the video “horse_7”. It includes one superset representing all horses within the scene and two subsets that specifically identify horses on the ground and horses in the river.

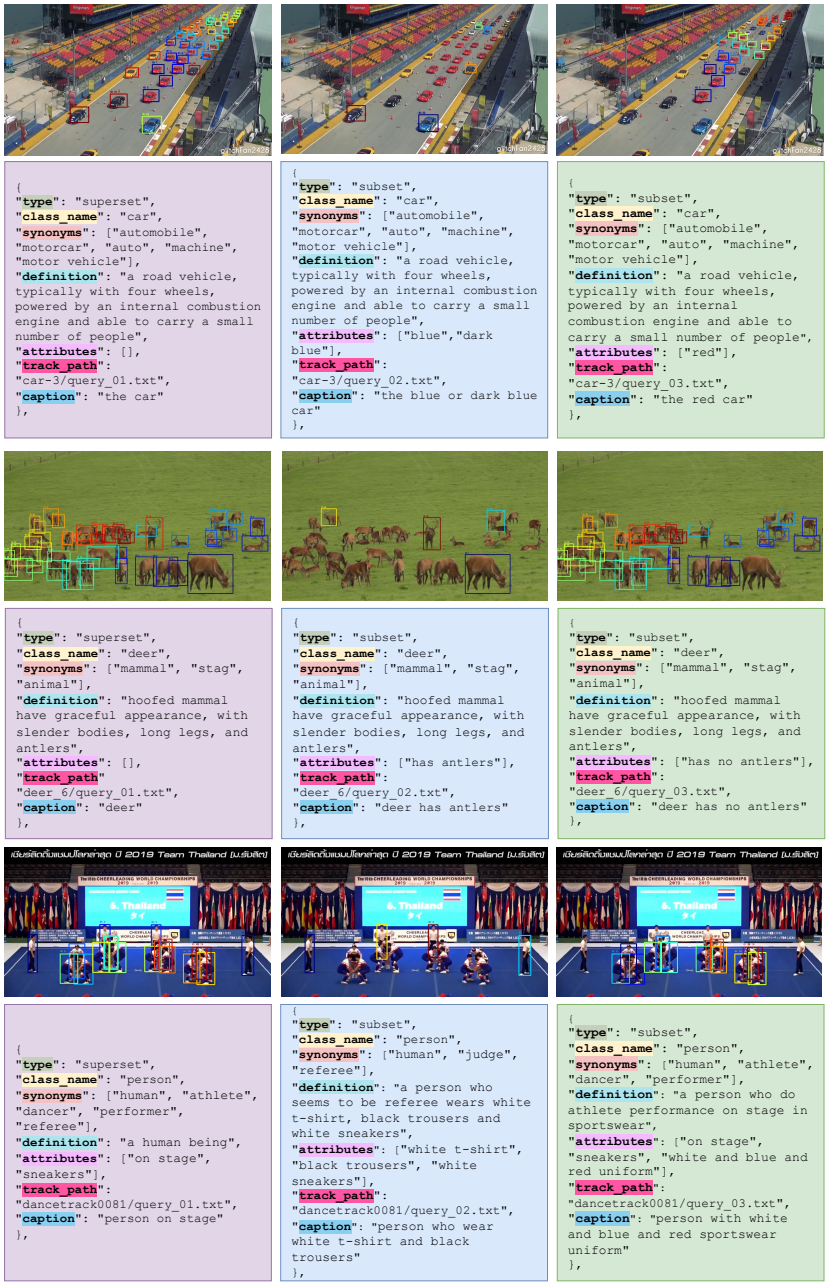


Fig. II: Additional examples of textual description provided in our G^{2}MOT.

G^2MOT infor	GMOT-40	AnimalTrack	DanceTrack	SportsMOT	G^2MOT
# superset queries	40	58	65	90	253
# subset queries	46	27	13	266	352
Textual description infor					
# captions	86	85	78	356	605
# definitions	86	85	78	356	605
# synonyms	350	489	382	1068	2289
Avg caption length	3.26	1.55	4.69	6.93	5.37
Avg definition length	14.53	14.0	7.88	4.00	7.71
Avg definition length	4.06	5.75	4.89	3.00	3.78
# words in caption	279	132	366	2470	3247
# words in definition	1249	1190	615	1424	4478

Fig. III: Statistical information from each sub-dataset (GMOT-40, AnimalTrack, DanceTrack, SportMOT) for the construction of our G^{2}MOT dataset.

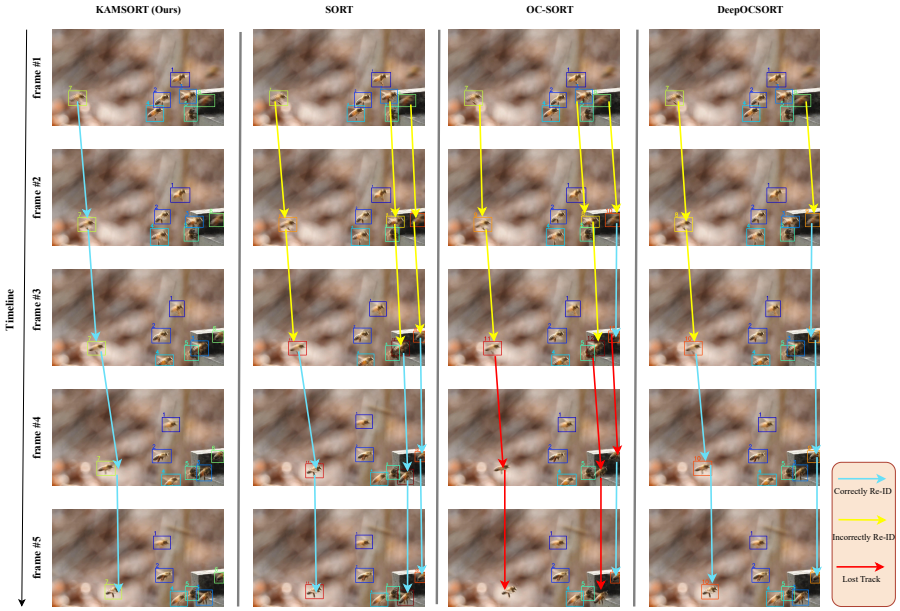


Fig. IV: Tracking comparison on *fast motion* objects between our KAM-SORT with SORT, OC-SORT and DeepOCSORT on the video "insect-3".

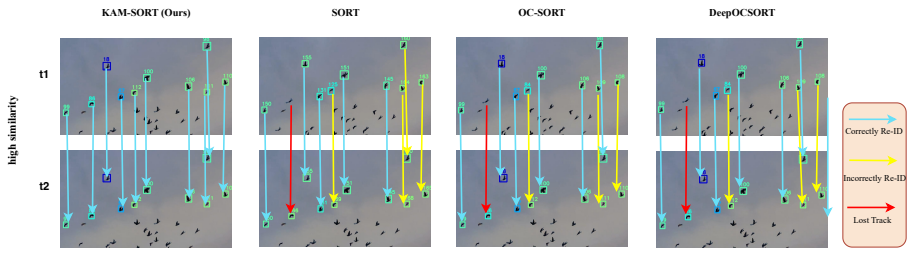


Fig. V: Tracking comparison on *high similarity in appearance* objects between our KAM-SORT with SORT, OC-SORT and DeepOCSORT on the video "bird-0".

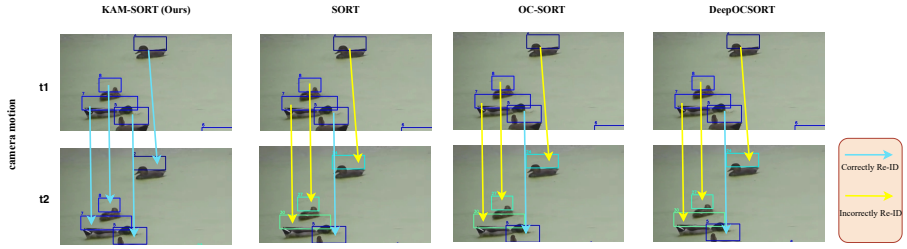


Fig. VI: Tracking comparison when both camera and object moving between our KAM-SORT with SORT, OC-SORT and DeepOCSORT on the video "duck-4".

References

1. 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)
2. 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)
3. 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)
4. 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)
5. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
6. 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)
7. 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)
8. Wu, D., Han, W., Wang, T., Dong, X., Zhang, X., Shen, J.: Referring multi-object tracking. In: CVPR. pp. 14633–14642 (2023)
9. Zhang, L., Gao, J., Xiao, Z., Fan, H.: Animaltrack: A benchmark for multi-animal tracking in the wild. IJCV pp. 1–18 (2022)