GMOT-40: A Benchmark for Generic Multiple Object Tracking

Hexin Bai[1]
Temple University
Philadelphia, USA
hexin.bai@temple.edu

Wensheng Cheng[1]
Stony Brook University
Stony Brook, USA
wenscheng@cs.stonybrook.edu

Peng Chu[1]
Microsoft
Redmond, USA
pengchu@microsoft.com

Juehuan Liu
Temple University
Philadelphia, USA
juehuan.liu@temple.edu

Kai Zhang
Temple University
Philadelphia, USA
zhang.kai@temple.edu

Haibin Ling[2]
Stony Brook University
Stony Brook, USA
hling@cs.stonybrook.edu

Abstract

Multiple Object Tracking (MOT) has witnessed remarkable advances in recent years. However, existing studies dominantly request prior knowledge of the tracking target (eg, pedestrians), and hence may not generalize well to unseen categories. In contrast, Generic Multiple Object Tracking (GMOT), which requires little prior information about the target, is largely under-explored. In this paper, we make contributions to boost the study of GMOT in three aspects. First, we construct the first publicly available dense GMOT dataset, dubbed GMOT-40, which contains 40 carefully annotated sequences evenly distributed among 10 object categories. In addition, two tracking protocols are adopted to evaluate different characteristics of tracking algorithms. Second, by noting the lack of devoted tracking algorithms, we have designed a series of baseline GMOT algorithms. Third, we perform a thorough evaluations on GMOT-40, involving popular MOT algorithms (with necessary modifications) and the proposed baselines. The GMOT-40 benchmark is publicly available at https://github.com/Spritea/GMOT40.

1. Introduction

Multiple Object Tracking (MOT) has long been studied in the computer vision community [13, 39], due to its wide range of applications such as in robotics, surveillance, autonomous driving, cell tracking, *etc*. Remarkable advances have been made recently in MOT, partly due to the progress of major components such as detection, single object tracking, association, *etc*. Another driving force comes from the popularization of MOT benchmarks (*e.g*., [22, 33, 41, 52, 62]). Despite the achievement, previous studies in MOT mostly focus on a specific object category of interest (pedestrian, car, cell, *etc*.) and rely on models of such objects. For example, detectors of such objects are often pre-trained offline, and motion patterns for specific objects are sometimes utilized as well. It remains unclear how well existing MOT algorithms generalize to unseen objects and hence constrains the expansion of MOT to new applications, especially those with limited data for training object detectors.

By contrast, Generic Multiple Object Tracking (GMOT), which requests no prior knowledge of the objects to be tracked, aims to deal with these issues. Hence GMOT could be applied in video editing, animal behaviour analysis, and vision based object counting. Despite its wide applications, it is however seriously under-explored, except for some early investigations [37, 38]. Comparing the progress in GMOT with that in MOT, we see a clear lack of GMOT benchmark, and the absence of GMOT baselines with effective deep learning ingredients. Note that we follow the definition of GMOT in [38], *i.e.*, tracking multiple objects of a generic object class.

Addressing the above issues, in this paper, we contribute to the study of GMOT in three aspects: dataset, baseline, and evaluation. First, we construct the first publicly available dense GMOT dataset, dubbed *GMOT-40*, for systematical study of GMOT. GMOT-40 contains 40 carefully selected sequences, which cover ten categories (*e.g*., *insect* and *balloon*) with four sequences per category. Each sequence contains multiple objects of same category, and the average number of objects per frame is around 22. All sequences are manually annotated with careful validation/correction. The sequences involve many challenging factors such as heavy blur, occlusion, *etc*. A tracking Proto-

[1]Equal contribution
[2]Corresponding author

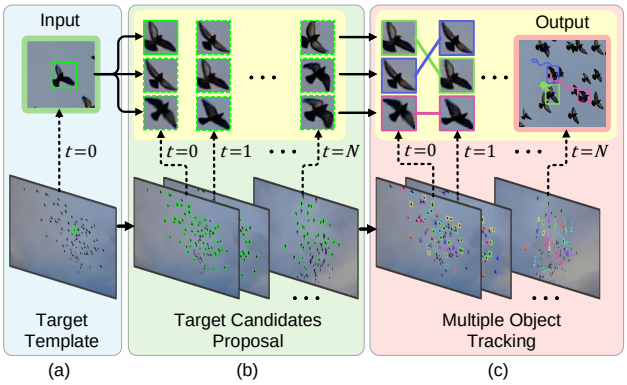

Figure 1. One-shot generic multiple object tracking (GMOT). (a): The input of one-shot generic MOT is a single bounding box to indicate a target template in the first frame. (b): The target template is used to discover and propose all other target candidates of same category, which is different than model-based MOT where a pre-trained detector (typically class-specific) is required. (c): MOT then can be performed on the proposed candidates in either an online or offline manner. Yellow rectangles are zoomed-in local views of targets.

col is adopted to evaluate different characteristics of tracking algorithms. The *one-shot GMOT* [37, 38], takes as input the bounding box of *one* target object in the first frame, and aims to detect and track *all* objects of the same category. Figure 1 illustrates the one-shot GMOT Protocol.

Second, we design a series of baseline tracking algorithms dedicated to one-shot GMOT. These baselines consist of a one-shot detection stage and a target association stage. The one-shot detection stage is adapted from the recently proposed GlobalTrack algorithm [28]. The target association stage comes from several typical MOT algorithms. For each baseline, the one-shot detection algorithm plays the role of public detector.

Third, we conduct thorough evaluations on GMOT-40. The evaluation involves both classic tracking algorithms (*e.g.*, [8, 53, 54]) and recently proposed one (*e.g.*, [12]), with necessary modifications. The results show that, as an important tracking problem, GMOT has a large room for improvement.

To summarize, we make three contributions in this paper:

- the first publicly available dense GMOT dataset, GMOT-40, which is carefully designed and annotated, along with evaluation Protocol,
- a series of GMOT baselines adapted from modern deep-learning enhanced MOT algorithm, and
- thorough evaluations and analysis on GMOT-40.

Table 1. Comparison of densely annotated data used in GMOT studies. # seq: number of sequence, # cat: number of categories, # tgt: average number of targets per frame. *: Estimated from samples in the paper.

Publication	Year	# seq.	# cat.	# tgt.
Luo *et al*. [37]	2013	4	4	$\approx 15^\star$
Zhang *et al*. [59]	2014	9	9	$\approx 3^\star$
Luo *et al*. [38]	2014	8	8	$\approx 15^\star$
Zhu *et al*. [61]	2017	3	1	13.13
Liu *et al*. [36]	2020	24	9	3.375
GMOT-40	2021	**40**	**10**	**26.58**

2. Related Work

2.1. MOT Algorithms

Multiple object tracking (MOT) has been an active research area for decades [13, 39]. Based on whether the target priors are presumed to the tracker, MOT approaches can be roughly categorized as model-based and model-free methods. In the context of model-based methods, the most popular framework is the tracking-by-detection one where a category-aware detector is employed for generating candidate proposals, and the tracker itself primarily focuses on solving the data association problem. Many methods have been investigated under this framework, such as Hungarian algorithm [6, 19, 26], network flow [16, 56, 58], graph multicut [25, 30, 50], multiple hypotheses tracking [11, 32] and multi-dimensional assignment [14, 47] using a variety of affinity estimation schemes. With recent advances in deep learning, deep neural networks are also learned to solve the data association problem [10, 12, 42].

Model-based MOT methods can automatically handle the entering and exiting events of targets. However, it heavily depends on using target priors by employing a category detector or the Re-identification (ReID) based affinity estimator. Therefore, most recent MOT methods in this category focus on pedestrian and vehicle tracking. For example, there is an increasing popularity in the community to leverage ReID dataset [34, 45, 60] or pose estimation dataset [2] to improve association robustness during tracking [10, 24, 29, 57], while others adopt the state-of-the-art person detection techniques, such as [3, 23, 43, 44, 46]. These detection and ReID networks are trained and hence limited by the available datasets, therefore, the generic targets will not be handled and tracked successfully by methods in this category.

Despite the dominant effort on the person and vehicle tracking, there are a number of works that have focused on other target categories. Cell tracking [7, 40, 51, 55] is a popular topic in this section. Detecting and tracking multiple objects, such as ants [31], bats [5], birds [38], bees [9] and fish [21, 48, 49] are also investigated. Methods proposed in those works also need special modeling of target appear-

ance or motion pattern thus cannot be applied generally in generic targets either.

Model-free methods contribute another category of solutions to MOT. Tracking without target prior is primarily proposed for solving *Single Object Tracking* (SOT) where only one bounding box of target is given at the first frame and no category prior is known to the tracker. It is an emerging topic to extend the model-free idea to the context of MOT. However there is no unified framework so far. In [59], structure information is used to help the tracking of multiple appearance-wise similar objects. Appearance and motion models are learned in [36] to tackle sudden appearance change and occlusion. Both the two methods need the manual initialization of all targets. In [61], a generic category independent object proposal module is used to generate target candidates. Luo *et al*. [38] proposed to use clustered Multiple Task Learning for generic object detection. All these works are evaluated on datasets that either have limited number of sequences or limited number of target categories.

2.2. MOT Benchmarks

There are multiple benchmark datasets for model-based MOT. One of the oldest benchmarks is the PETS benchmark [20] which contains three sequences for single camera MOT while all of them are on pedestrians. Later on, a benchmark mainly for autonomous driving is KITTI [22] which contains two categories of pedestrian and vehicle. After that, a benchmark dataset solely on pedestrian tracking was proposed by Alahi *et al*. [1]. Although this benchmark contains 42 million pedestrian trajectories, yet its annotation is not high-quality (*i.e*., not annotated by human). Then a MOT benchmark dataset on vehicle tracking was released with the name UA-DETRAC [52] which contains 100 sequences. In the same year MOT15 was released [33] which organized the publicly available MOT data by then and became one of the most popular MOT benchmarks. Yet it is worth noting that there are just two categories: people and vehicle in this benchmark, and only 22 sequences are included. Later, MOT16 [41] was published with 14 sequences, devoted to people and vehicle tracking. VisDrone [62] was released with 96 sequences focused on vehicle and people.

In addition to the popular MOT benchmark dataset mentioned above on people and vehicle tracking, there are some other benchmark datasets on special classes such as honey bees and cells. For example, the multiple cell tracking dataset [51] has 52 sequences with a focus on cell, the honey-bee tracking dataset [9] has 60 sequences of the honey bee.

As shown in Table 1, high quality datasets dedicated for model-free MOT are rare. In [59], Zhang *et al*. collect a dataset with nine video sequences, each for a different type of target. Among the videos, three are adapted from a SOT dataset, while the rest videos are collected from YouTube. The dataset contains average of 3 targets per frame. Each video here has average of 842 frames in length. Targets in the dataset are present all-time in the video, which relieves the tracker of handling the entering and exiting event of targets. Luo *et al*. collected datasets with four and eight videos in [37] and [38] respectively for an early study of GMOT. Recent works [61, 36] tend to use mixed sequences picked from other SOT or multiple pedestrian tracking datasets. Recently, a large-scale benchmark for tracking any object (TAO) is proposed [15]. However, TAO is not densely annotated and has low annotation quality. Only one out of every 30 frames is annotated by hand, and the average trajectories of TAO in each sequence is only 5.9. Besides, the task of TAO is to track multiple objects of different classes, which differs with the GMOT concept in this paper. Hence we do not include TAO in comparison Table 1.

Compared with the data used in previous studies, our proposed GMOT-40 dataset provides the the first publicly available dense dataset on GMOT. GMOT-40 contains more sequences and categories than previous GMOT datasets. Moreover, the target density in GMOT-40 is much higher than existing datasets, *e.g*., 26.58 per sequence *vs* 5.9 per sequence in TAO, and the sequences involve many real-world challenges such as entering and exiting events, fast motion, occlusion, *etc*. As a result, the release of GMOT-40 is expected to largely facilitate future research in GMOT.

3. The Generic MOT Dataset GMOT-40

In this section, we will present the GMOT-40 dataset and the associated evaluation protocol. As described in the related work, a serious GMOT dataset/benchmark is in great need for advancing the study of GMOT. By investigating the data issues in previous papers and borrowing ideas from recently popularized tracking benchmarks, we aim to construct a high-quality dataset in the following aspects:

- *Diversity in target category.* To address the generalization concern in previous MOT studies, GMOT-40 is designed to contain 40 sequences from 10 different categories, which is larger than most of previously studied datasets (typically less than 3 categories). The four sequences in each category are designed with further diversity. For example, the "person" category in GMOT-40 covers both normal "person" as in PASCAL-VOC [17] and an unseen type "wingsuit"; the "insect" category covers "ant" and "bee", both of which are unseen in MS-COCO [35] or PASCAL-VOCC [17]. Some sample frames in GMOT-40 are shown in Figure 2.

- *Real world challenges.* During sequence selection, we pay special attention to include sequences with vari-

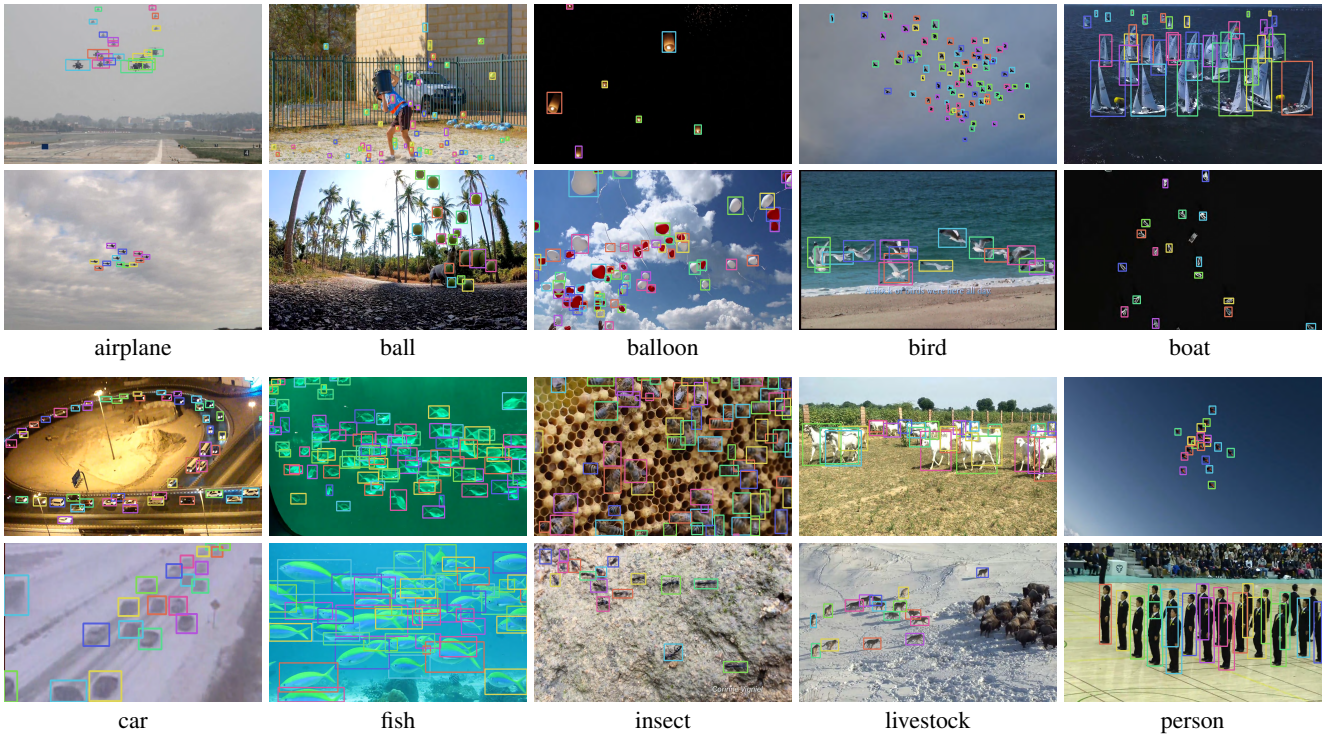

Figure 2. Samples from each category of GMOT-40.

ous real-world challenges such as occlusion, target enter/exiting, fast motion, blur, *etc*. Moreover, the target density ranges from 3 to 100 targets per frame, with the average around 26. All these properties make GMOT-40 cover a wide range of scenarios.

- *High-quality annotation.* For high quality annotation, each frame in the sequence should be annotated by hand to ensure precise annotation. Besides, the initial annotation will be followed by careful validation and revision.

It is worth noting that, while more sequences would likely further improve the data usability, the additional non-trivial efforts in manual annotation may postpone the timely release of the dataset. In fact, as shown in Table 1, GMOT-40 brings comprehensive improvements over previously used GMOT data, and is thus expected to facilitate the GMOT research in the future.

3.1. Data Collection

With the guidance mentioned above, we start by deciding 10 categories of objects that are highly possible to be dense and crowded. When selecting video sequences, we request that at least 80% of the frames in a sequence to have more than 10 targets. Most targets of same category have similar appearance, while part of them differs on appearance, which is more close to reality. The minimum length of the sequence is set to 100 frames.

After classes and requirements are determined, we started searching the YouTube with possible candidate videos. About 1000 sequences are initially picked as candidates. After scrutiny, we select 40 sequences out of them for better quality and more challenging task. Yet it does not mean that these 40 sequences are ready for annotation. Some of the sequences contain a large part that is irrelevant to our task. For example, in "balloon" category, there are starting and ending sections focusing on the stage or the crowd of the celebration in the festival, which should be removed. In such a way, we carefully edit the video and select the best clips with a minimum of 100 frames.

Finally, GMOT-40 contains 50.65 trajectories per sequence on average. The whole dataset includes 9,643 frames in total, and each sequence has an average length of 240 frames. 85.28% of the frames have more than 10 targets. The FPS ranges from 24 to 30 while resolution ranges from 480p to 1080p.

The statistics of GMOT-40 in comparison with other densely annotated data used in GMOT studies are summarized in Table 1. Note that we use the category definition of GMOT-40 here, since categories in other benchmarks are not general enough. As an example, both "sky diving" and "basketball" classes in [36] belong to the "person" class of GMOT-40.

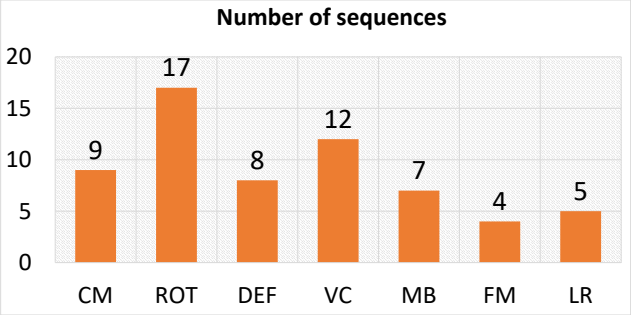

Figure 3. Number of sequences for different attributes in our GMOT-40.

3.2. Annotation

The annotation format follows that of MOT15 [33] where the detailed description is in the Supplementary Material. The only difference is that there is no out-of-view value and hence all bounding box in the groundtruth file should be considered in evaluation protocols.

Furthermore, only targets in the same category are annotated. For example, only the wolf in the "stock" category is annotated as shown in Figure 2 since the initial bounding box indicates that only the wolf is the object of interest. Besides, the targets in the same categories are treated indiscriminately such as the red and white balloons in Figure 2.

The most important parts for building a high-quality GMOT dataset are manual labeling, double-checking, and error-correction. To ensure this, a group of experts such as Ph.D. students are included in the annotation team. For each video, it is first sent to the labeler to decide the group of interest. Then an expert will review the target group to see whether it reaches our requirement. After approval by experts, the labeler will start working on the annotation. The completed annotation will again be sent to experts for review and possible revision.

3.3. Video Attributes

As shown in the Figure 2, diverse scenarios and hence more comprehensive attributes are included in GMOT-40 compared with other data used in previous GMOT papers. As an example, all of the "person", "ball" and "insect" classes have the properties of motion-blur and fast motion. Besides, the viewpoint significantly affects the appearance in "boat" category. Furthermore, low resolution and camera motion appear in "ball" and "livestock" respectively.

A detailed histogram on various attributes are presented in Figure 3. The abbreviation of attributes have the following meaning: **CM** – camera motion; **ROT** – target rotation; **DEF** – target deforms in the tracking; **VC** – significant viewpoint change that affects the appearance of target; **MB** – target is blurred due to camera or target motion; **FM** – fast motion of the targets with displacements larger than the bounding box; **LR** – target bounding box is smaller than 1024 pixel for at least 30% of the targets in the whole sequences.

Although some of the attributes above are present in previous studies of GMOT [36, 37, 38, 59, 61], yet GMOT-40 is the most comprehensive one, since it is collected from various natural scenes. These miscellaneous attributes of GMOT-40 can help the community to evaluate their trackers from multiple aspects.

4. GMOT Protocols and Tracking Baselines

4.1. Protocol

Associated with the GMOT-40 dataset, we design a dedicated one-shot evaluation protocol for GMOT, adapting the settings from previous works such as in [38]. To facilitate the developing of GMOT trackers, an ablation study is also implemented to evaluate the association ability of tracker.

This protocol is to comprehensively evaluate the GMOT trackers in real-world application settings. As claimed in [38], a practical generic tracker is model-free thus is able to track multiple generic objects knowing only one template of targets. By adopting this Protocol, only one bounding box in the first frame of each video is provided to indicate the objects of interest. Trackers are supposed to use the object in that bounding box as a template and leverage the information of that object to detect and track all the targets in the video of same category. All sequences in GMOT-40 are used to test the tracker for their performance on unseen category for the one-shot GMOT protocol. For comparison, we also design several new baselines (see Section 4.2) to generate the public detection for the whole sequence, using the only one sample given in the first frame. Trackers can be trained at any other benchmarks except GMOT-40.

To choose the initial target of one sequence, we randomly sample some targets in the first frame that are not occluded. Then we carefully pick the best one out of them by hand to ensure it is representative and robust as the one-shot sample.

4.2. Baselines for One-shot GMOT

For one-shot GMOT protocol, we propose a series of two-stage baselines by adapting existing tracking algorithms. Each baseline consists of an one-shot detection stage, which gets detection results for all frames in sequence, and a target association stage, which associates detected targets and gets the final tracking results.

4.2.1 One-Shot Detection Stage

In our implementation, we adopt a recently proposed SOT method, GlobalTrack [28], to create a one-shot detection

method. GlobalTrack searches the whole image in following frames (search frames) while most SOT trackers only search a predefined neighborhood of the target position in the previous frame. The model is pretrained on other datasets [35, 27, 18]. We then split the modified model to two modules, a target-guided region proposal module, and a target-guided matching module. The target-guided region proposal module extracts features for the labeled target on the initial frame, and return regions that may contain targets on the search frame. Then target-guided matching module extracts features from these regions, computes similarity scores between these potential targets, and produces multiple search results with the refined position. Furthermore, those targets with similarity scores lower than the threshold (0.1) are filtered out.

In the one-shot detection process, the initial frame is always the first frame and the search frames include all frames in the sequence, including the first frame itself. The detection process is repeated to get results for all these frames. The whole process is shown in Algorithm 1.

4.2.2 Target Association Stage

With these detection results, we now transform the one-shot GMOT task to a traditional MOT task with public detection. Most existing MOT algorithms can be adapted here to get association. The MOT algorithms used in evaluation are stated in Section 5.2.

Combining the one-shot detection method with different target association methods, we get a series of baselines for the one-shot GMOT task. We evaluate their tracking performances comprehensively in Section 5.3.

Algorithm 1: One-shot Detection Process.

Data:
$\{I_1, ..., I_m\}$: images in a sequence;
x_{gt}: initial detection (groundtruth box) in I_1;
s_{th}: threshold for detection similarity score.
Model:
ϕ_R: target-guided region proposal module;
ϕ_M: target-guided matching module.
Output:
$\{x_i^k\}_{i=1}^{n_k}$: n_k detected targets for I_k, $1 \leq k \leq m$.

Extract features for the initial target;
    $F_{gt} = \phi_R(I_1, x_{gt})$;
for $k = 1, ..., m$ **do**
    Use F_{gt}, ϕ_R to produce r_k regions R that may contain targets on image I_k;
        $R = \{x_1^k, ..., x_{r_k}^k\} = \phi_R(F_{gt}, I_k)$;
    Use ϕ_M to extract features F_R from R;
        $F_R = \{f_1^k, ..., f_{r_k}^k\} = \phi_M(R)$;
    Compute similarity scores S between F_R and F_{gt}, and produce targets T with refined positions;
        $S = \{s_1^k, ..., s_{r_k}^k\} = \phi_M(F_{gt}, F_R)$;
        $T = \{\tilde{x}_1^k, ..., \tilde{x}_{r_k}^k\} = \phi_M(F_{gt}, F_R)$;
    Filter T by comparing S with s_{th}, and then get the final n_k targets T^k;
        $T^k = \{x_1^k, ..., x_{n_k}^k\} = C(T, S, s_{th})$;
    where C denotes the comparison process;
end

5. Experiment

5.1. Evaluation Metrics

A group of metrics on MOT has been proposed to fairly compare the tracker and reveal the performance. Among them the most widely used ones are CLEAR MOT metrics [4] and ID metrics [45]. The former stresses the number of incorrect predictions while the latter focus on the longest time of following targets. Combining them will provide a comprehensive evaluation of the performance in GMOT-40.

5.2. Evaluated Trackers

We focus on the trackers that are built on public detection and have publicly available code. Both classical and more recent trackers are included to provide a comprehensive review. Among them, there are FAMNet [12], Deep SORT [53], MDP [54], IOU tracker [8].

5.3. Protocol Evaluation

We first evaluate the quality of the proposed target candidates that are generated by our baseline algorithm. Since in one-shot generic setting, the difference between categories is inconsequential. Thus we directly use AP (Average Precision) as our metric to report the "detection" solely performance. We have AP_{50} of 15.65% and AP_{75} of 15.51% while setting the IOU threshold at 0.5 and 0.75 respectively. Note that our baseline target candidate proposal is not trained on GMOT-40. In qualitative analysis, the baseline is found out to behave badly with deformation, rotation out-of-plane, motion blur and low resolution. The reason may be that the matching module of our modified GlobalTrack produced too many false negatives while ranking the confidence in the final stage.

The detection results generated by our baseline algorithm serve as public detection in the following experiments. We test the trackers on all 40 sequences in its ini-

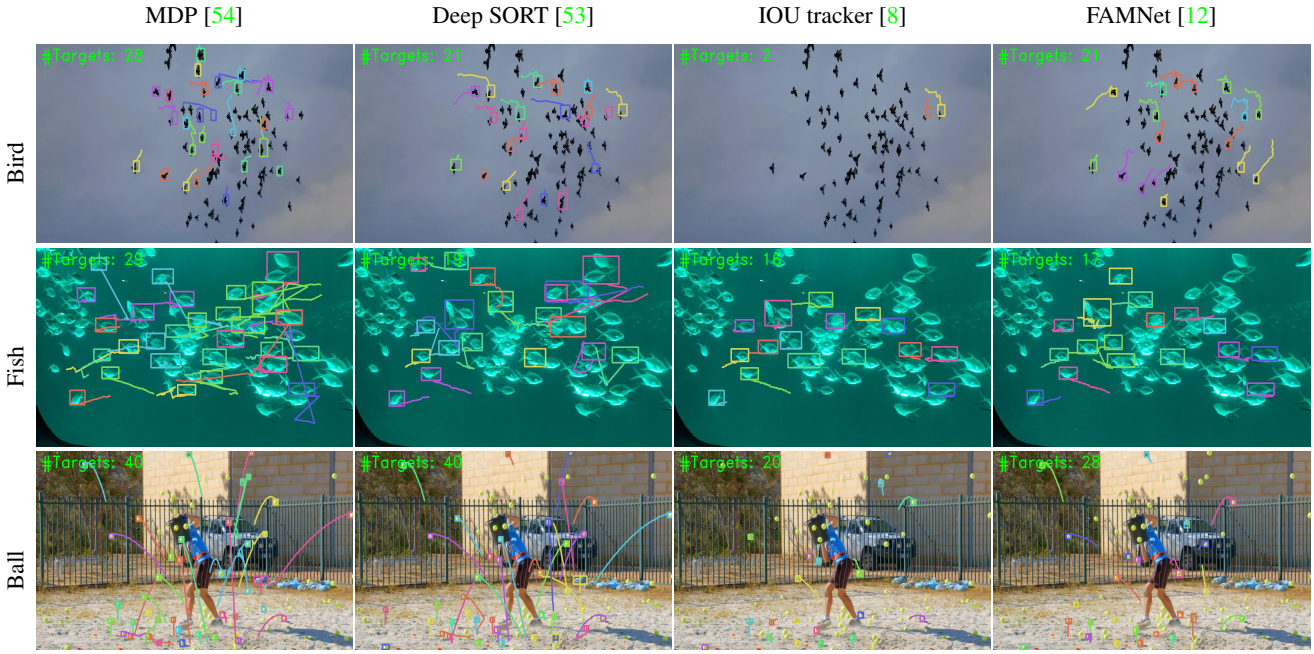

Figure 4. Results visualization of four trackers on sequences.

	MOTA	IDF1	IDP	IDR	Rcll	Prcn	MT	PT	ML↓	FP↓	FN↓	IDs↓	FM↓
MDP [54]	**19.80%**	**31.30%**	61.80%	**21.00%**	**27.20%**	80.20%	142	**621**	**1161**	17260	**186580**	1779	2748
DeepSORT [53]	14.50%	24.40%	**67.50%**	14.90%	18.50%	**84.10%**	72	509	1363	9000	208818	1315	2233
IOU [8]	11.80%	20.30%	64.60%	12.00%	15.40%	82.60%	56	397	1491	**8299**	216921	**754**	**1668**
FAMNet [12]	18.00%	28.30%	54.80%	19.10%	26.80%	76.80%	**166**	581	1197	20741	187730	1660	1878

Table 2. Comparison of trackers with one-shot GMOT protocol.

Methods	MOTA	MOTP	IDF1
MDP[54]	19.92% ± 1.84%	24.16% ± 0.27%	31.84% ± 2.23%
DeepSORT[53]	14.98% ± 1.47%	23.66% ± 0.53%	25.38% ± 2.32%
IOU[8]	12.36% ± 1.60%	25.34% ± 0.36%	20.90% ± 2.73%
FAMNet[12]	17.60% ± 0.85%	22.56% ± 0.23%	27.76% ± 1.16%

Table 3. Average of five runs initiated by randomly picked one-shot templates.

tial setting with the pre-trained model without any further modification. The results as well as MOTA and IDF1 are listed in the Table 2. With the inclusion of the one-shot detector, MDP becomes the best among them all. Yet its IDF1 is just 31.30% and MOTA is just 19.80% . Deep SORT and FAMNet here behave slightly worse than MDP with the IOU tracker after them. In other words, there is correlation between their processing of detection and their performance. A sample of results is presented Figure 4 with each color standing for a different trajectory.

Besides, we include Figure 5 to compare the performance in different classes. Each bar represents the mean of all 5 trackers. Specifically, the "bird" and "insect" classes poses a challenge for all the trackers. This again proves the necessity of diversity and hence the release of GMOT-40. A more detailed version is included in Supplementary Material.

Finally, to make sure the results in experiment is unbiased from the initial results picked by user. We randomly sample the one target in the 1st frame for protocol and repeat this procedure for 5 times. Then we report the mean and standard deviation of the results over these 5 experiments. The results are shown in Table 3. As we can see, the fluctuations are very low, implying that the choice of the initial bounding box does not affect the result significantly.

5.4. Ablation Study

In ablation study, the groundtruth detection are provided for the tracker while all other experiment conditions are the same. The result of this protocol is presented in Table 4, where we can see nearly all trackers' performances improve significantly compared with Table 2. Note that our benchmark contains many categories that are unseen for the tracker during their training. Hence the benchmark would favor the association based on Intersection Over Union (IOU) of targets across frames rather than appearance fea-

	MOTA	IDF1	IDP	IDR	Rcll	Prcn	MT	PT	ML↓	FP↓	FN↓	IDs↓	FM↓
MDP [54]	75.00%	72.50%	79.50%	66.70%	80.70%	96.20%	1105	**703**	**136**	8234	49448	4103	4758
DeepSORT [53]	**80.60%**	**79.30%**	85.30%	**74.00%**	**84.50%**	97.30%	**1344**	344	256	5944	**39648**	4074	**2937**
IOU [8]	75.90%	79.00%	85.80%	73.20%	80.40%	94.20%	1237	260	447	12704	50232	**1225**	3767
FAMNet [12]	67.40%	70.50%	**86.30%**	60.50%	70.10%	**97.60%**	1302	319	323	**4505**	76706	2454	6229

Table 4. Comparison of trackers with the protocol in ablation study.

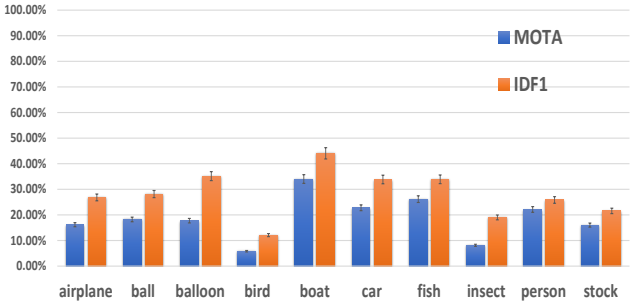

Figure 5. Average scores of all trackers for different classes in one-shot GMOT Protocol.

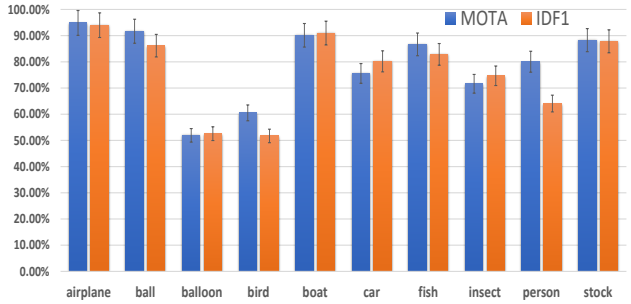

Figure 6. Average scores of all trackers for different classes with the protocol in ablation study.

tures. As a result, the simple IOU tracker has the 2nd best IDF1 and MOTA of 79.00% and 75.90%, respectively. While using both motion and appearance information, Deep SORT has the best MOTA and IDF1 score by maintaining a reasonable balance between them. For MDP, its performance is not as good as Deep SORT and IOU tracker. The reason may be its superfluous processing on detection since we directly provide groundtruth detection here. For FAMNet [12], its mediocre performance is mainly due to processing on detection noise. Although groundtruth detection are provided here, FAMNet drops too many detection and hence causes many false negatives.

Furthermore, we include Figure 6 to compare the performance under different categories. Generally speaking, the trackers perform much better in ablation study. The difference in performance among categories emphasizes the importance of releasing a GMOT benchmark to evaluate trackers more comprehensively.

6. Conclusion

In this paper, we proposed the first, to the best of our knowledge, publicly available densely annotated generic multiple object tracking (GMOT) benchmark named GMOT-40. By thoroughly considering major MOT factors and carefully annotating all tracking objects, GMOT-40 contains 40 sequences evenly distributed among 10 object categories. Associated with the GMOT-40 dataset is the one-shot evaluation protocol for GMOT. Several new baseline algorithms dedicated to one-shot GMOT are developed as well, and evaluated together with relevant MOT trackers to provide references for future study. The evaluation shows that there is still large room to improve for GMOT and further studies are desired. Overall, we expect the benchmark, along with the initial studies, to largely facilitate future research on GMOT, which is an important yet under-explored problem in computer vision.

Acknowledgements. We thank the anonymous reviewers for their insightful suggestions that largely help improve the work. This work was supported in part by US National Science Foundation (No. 1814745 and No. 2006665).

References

[1] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *CVPR*, 2014. 3

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 2

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, 2019. 2

[4] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008. 6

[5] Margrit Betke, Diane E Hirsh, Angshuman Bagchi, Nickolay I Hristov, Nicholas C Makris, and Thomas H Kunz. Tracking large variable numbers of objects in clutter. In *CVPR*, 2007. 2

[6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, 2016. 2

[7] Ryoma Bise, Zhaozheng Yin, and Takeo Kanade. Reliable cell tracking by global data association. In *ISBI*, 2011. 2

[8] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *AVSS*, 2017. 2, 6, 7, 8

[9] Katarzyna Bozek, Laetitia Hebert, Alexander S Mikheyev, and Greg J Stephens. Towards dense object tracking in a 2d honeybee hive. In *CVPR*, 2018. 2, 3

[10] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *CVPR*, 2020. 2

[11] Jiahui Chen, Hao Sheng, Yang Zhang, and Zhang Xiong. Enhancing detection model for multiple hypothesis tracking. In *CVPRW*, 2017. 2

[12] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *ICCV*, 2019. 2, 6, 7, 8

[13] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 2020. 1, 2

[14] Robert T Collins. Multitarget data association with higher-order motion models. In *CVPR*, 2012. 2

[15] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *ECCV*, pages 436–454. Springer, 2020. 3

[16] Afshin Dehghan, Yicong Tian, Philip HS Torr, and Mubarak Shah. Target identity-aware network flow for online multiple target tracking. In *CVPR*, 2015. 2

[17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 3

[18] Heng Fan, Hexin Bai, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Harshit, Mingzhen Huang, Juehuan Liu, et al. LaSOT: A high-quality large-scale single object tracking benchmark. *IJCV*, 2020. 6

[19] Kuan Fang, Yu Xiang, Xiaocheng Li, and Silvio Savarese. Recurrent autoregressive networks for online multi-object tracking. In *WACV*, 2018. 2

[20] James Ferryman and Ali Shahrokni. Pets2009: Dataset and challenge. In *PETS*, 2009. 3

[21] Ebraheem Fontaine, Alan H Barr, and Joel W Burdick. Model-based tracking of multiple worms and fish. In *ICCVW*, 2008. 2

[22] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 3

[23] Shoudong Han, Piao Huang, Hongwei Wang, En Yu, Donghaisheng Liu, Xiaofeng Pan, and Jun Zhao. MAT: Motion-aware multi-object tracking. *arXiv preprint arXiv:2009.04794*, 2020. 2

[24] Roberto Henschel, Yunzhe Zou, and Bodo Rosenhahn. Multiple people tracking using body and joint detections. In *CVPRW*, 2019. 2

[25] Andrea Hornakova, Roberto Henschel, Bodo Rosenhahn, and Paul Swoboda. Lifted disjoint paths with application in multiple object tracking. In *ICML*, 2020. 2

[26] Chang Huang, Bo Wu, and Ramakant Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008. 2

[27] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *TPAMI*, 2019. 6

[28] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. *AAAI*, 2020. 2, 5

[29] Shyamgopal Karthik, Ameya Prabhu, and Vineet Gandhi. Simple unsupervised multi-object tracking. *arXiv preprint arXiv:2006.02609*, 2020. 2

[30] Margret Keuper, Evgeny Levinkov, Nicolas Bonneel, Guillaume Lavoué, Thomas Brox, and Bjorn Andres. Efficient decomposition of image and mesh graphs by lifted multicuts. In *ICCV*, 2015. 2

[31] Zia Khan, Tucker Balch, and Frank Dellaert. An MCMC-based particle filter for tracking multiple interacting targets. In *ECCV*, 2004. 2

[32] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M Rehg. Multiple hypothesis tracking revisited. In *ICCV*, 2015. 2

[33] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942*, 2015. 1, 3, 5

[34] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3, 6

[36] Chongyu Liu, Rui Yao, S Hamid Rezatofighi, Ian Reid, and Qinfeng Shi. Model-free tracker for multiple objects using joint appearance and motion inference. *TIP*, 2020. 2, 3, 4, 5

[37] Wenhan Luo and Tae-Kyun Kim. Generic object crowd tracking by multi-task learning. In *BMVC*, 2013. 1, 2, 3, 5

[38] Wenhan Luo, Tae-Kyun Kim, Bjorn Stenger, Xiaowei Zhao, and Roberto Cipolla. Bi-label propagation for generic multiple object tracking. In *CVPR*, 2014. 1, 2, 3, 5

[39] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, Xiaowei Zhao, and Tae-Kyun Kim. Multiple object tracking: A literature review. *arXiv preprint arXiv:1409.7618*, 2014. 1, 2

[40] Martin Maška, Vladimír Ulman, David Svoboda, Pavel Matula, Petr Matula, Cristina Ederra, Ainhoa Urbiola, Tomás España, Subramanian Venkatesan, Deepak MW Balak, et al. A benchmark for comparison of cell tracking algorithms. *Bioinformatics*, 2014. 2

[41] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831*, 2016. 1, 3

[42] Anton Milan, Seyed Hamid Rezatofighi, Anthony R Dick, Ian D Reid, and Konrad Schindler. Online multi-target tracking using recurrent neural networks. In *AAAI*, 2017. 2

[43] Bo Pang, Yizhuo Li, Yifan Zhang, Muchen Li, and Cewu Lu. TubeTK: Adopting tubes to track multi-object in a one-step training model. In *CVPR*, 2020. 2

[44] Jinlong Peng, Changan Wang, Fangbin Wan, Yang Wu, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *ECCV*, 2020. 2

[45] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016. 2, 6

[46] Chaobing Shan, Chunbo Wei, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Xiaoliang Cheng, and Kewei Liang. FGAGT: Flow-guided adaptive graph tracking. *arXiv preprint arXiv:2010.09015*, 2020. 2

[47] Xinchu Shi, Haibin Ling, Yu Pang, Weiming Hu, Peng Chu, and Junliang Xing. Rank-1 tensor approximation for high-order association in multi-target tracking. *IJCV*, 2019. 2

[48] Concetto Spampinato, Yun-Heh Chen-Burger, Gayathri Nadarajan, and Robert B Fisher. Detecting, tracking and counting fish in low quality unconstrained underwater videos. In *VISAPP*, 2008. 2

[49] Concetto Spampinato, Simone Palazzo, Daniela Giordano, Isaak Kavasidis, Fang-Pang Lin, and Yun-Te Lin. Covariance based fish tracking in real-life underwater environment. In *VISAPP*, 2012. 2

[50] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *CVPR*, 2017. 2

[51] Vladimír Ulman, Martin Maška, Klas EG Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, et al. An objective comparison of cell-tracking algorithms. *Nature methods*, 2017. 2, 3

[52] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *CVIU*, 2020. 1, 3

[53] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 2, 6, 7, 8

[54] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *ICCV*, 2015. 2, 6, 7, 8

[55] Fan Yang, Fariborz Soroush, Ge Deng, Sijia Yu, Peng Chu, Mohammad F Kiani, and Haibin Ling. Multiple neutrophils tracking in vitro array using high-order temporal information. In *EMBC*, 2018. 2

[56] Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *ECCV*, 2012. 2

[57] Yifu Zhan, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. A simple baseline for multi-object tracking. *arXiv preprint arXiv:2004.01888*, 2020. 2

[58] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008. 2

[59] Lu Zhang and Laurens Van Der Maaten. Preserving structure in model-free tracking. *TPAMI*, 2013. 2, 3, 5

[60] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2

[61] Gao Zhu, Fatih Porikli, and Hongdong Li. Model-free multiple object tracking with shared proposals. In *ACCV*, 2016. 2, 3, 5

[62] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Qinghua Hu, and Haibin Ling. Vision meets drones: Past, present and future. *arXiv preprint arXiv:2001.06303*, 2020. 1, 3