GMOT-40: A Benchmark for Generic Multiple Object Tracking

Hexin Bai
temple University
Philadelphia, USA
hexin.bai@temple.edu

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

Peng Chu
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
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 (e.g., 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 evaluation 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 systematic 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-
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 the same category, which is different than model-based MOT where a pretrained 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.

The evaluation involves both classic tracking 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.

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-

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.

<table>
<thead>
<tr>
<th>Publication</th>
<th>Year</th>
<th># seq.</th>
<th># cat.</th>
<th># tgt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luo et al. [37]</td>
<td>2013</td>
<td>4</td>
<td>4</td>
<td>≈15*</td>
</tr>
<tr>
<td>Zhang et al. [59]</td>
<td>2014</td>
<td>9</td>
<td>9</td>
<td>≈3*</td>
</tr>
<tr>
<td>Luo et al. [38]</td>
<td>2014</td>
<td>8</td>
<td>8</td>
<td>≈15*</td>
</tr>
<tr>
<td>Zhu et al. [61]</td>
<td>2017</td>
<td>3</td>
<td>1</td>
<td>13.13</td>
</tr>
<tr>
<td>Liu et al. [36]</td>
<td>2020</td>
<td>24</td>
<td>9</td>
<td>3,375</td>
</tr>
<tr>
<td>GMOT-40</td>
<td>2021</td>
<td>40</td>
<td>10</td>
<td>26.58</td>
</tr>
</tbody>
</table>
ance or motion pattern thus cannot be applied generally in
generic targets either.

Model-free methods contribute another category of so-
lutions 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 emerg-
ing 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 multi-
ple appearance-wise similar objects. Appearance and mo-
tion models are learned in [36] to tackle sudden appearance
change and occlusion. Both the two methods need the man-
ual initialization of all targets. In [61], a generic category
independent object proposal module is used to generate tar-
get 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 lim-
ited number of sequences or limited number of target cate-
gories.

2.2. MOT Benchmarks

There are multiple benchmark datasets for model-based
MOT. One of the oldest benchmarks is the PETS bench-
mark [20] which contains three sequences for single cam-
era 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 track-
ing was proposed by Alahi et al. [1]. Although this track-
mark contains 42 million pedestrian trajectories, yet its an-
notation 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: peo-
ple and vehicle in this benchmark, and only 22 sequences
are included. Later, MOT16 [41] was published with 14 se-
quencies, 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 men-
tioned 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 tar-
gets. 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 an-
notated and has low annotation quality. Only one out of
every 30 frames is annotated by hand, and the average tra-
jectories 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 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 ex-
pected 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 re-
lated 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 con-
struct a high-quality dataset in the following aspects:

- **Diversity in target category.** To address the gen-
  eralization concern in previous MOT studies, GMOT-40
  is designed to contain 40 sequences from 10 differ-
  ent 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” cate-
  gory 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

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

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>MOTA</th>
<th>MOTP</th>
<th>IDF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDP [54]</td>
<td>19.80%</td>
<td>31.30%</td>
<td>61.80%</td>
</tr>
<tr>
<td>DeepSORT [53]</td>
<td>14.50%</td>
<td>24.40%</td>
<td>67.50%</td>
</tr>
<tr>
<td>IOU [8]</td>
<td>11.80%</td>
<td>20.30%</td>
<td>64.60%</td>
</tr>
<tr>
<td>FAMNet [12]</td>
<td>18.00%</td>
<td>28.30%</td>
<td>54.80%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>MOTA</th>
<th>MOTP</th>
<th>IDF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDP [54]</td>
<td>19.92%</td>
<td>24.16%</td>
<td>31.84%</td>
</tr>
<tr>
<td>DeepSORT [53]</td>
<td>14.98%</td>
<td>23.66%</td>
<td>25.38%</td>
</tr>
<tr>
<td>IOU [8]</td>
<td>12.36%</td>
<td>25.34%</td>
<td>20.90%</td>
</tr>
<tr>
<td>FAMNet [12]</td>
<td>17.60%</td>
<td>22.56%</td>
<td>27.76%</td>
</tr>
</tbody>
</table>

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-
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, DeepSORT has the best MOTA and IDF1 score by maintaining a reasonable balance between them. For MDP, its performance is not as good as DeepSORT 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

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