# A Simple Baseline with Cascade Association for Robust Multi-Objects Tracking

Jiasheng Tang, Fei Du, Weihua Chen, Hao Luo, Fan Wang, Hao Li
Machine Intelligence Technology Lab, Alibaba Group
{jiasheng.tjs, dufei.df, kugang.cwh, michuan.lh, fan.w, lihao.lh}@alibaba-inc.com

## Abstract

*This report describes our solution to the challenge on the workshop of Robust Video Scene Understanding: Tracking and Video Segmentation (RobMOTS) at CVPR 2021. With given strong detections and segmentations, classical tracking-by-detection paradigm is used and a simple cascade association pipeline is built to track multiple objects robustly. Firstly, we extract feature embeddings of the detected objects as their appearance features. A general object Re-Identification (ReID) model is applied for feature extraction of all object categories except the class of "person" and "car". Thanks to the abundant public datasets in person ReID and vehicle ReID community, dedicated ReID models are trained for these two classes respectively. The appearance model for general objects is trained with 310,000 IDs from a collection of Single Object Tracking datasets, which is the same as the one used in the winning solution of the Tracking Any Object challenge associated with ECCV 2020. There are multiple steps in data association. The first step is based solely on the appearance features with a relatively strict threshold to guarantee the association precision; next, the remaining unmatched tracks and unmatched detections is linked by their IoU scores, similar to the provided STP baseline. After these two steps of data association on detection-level, reliable tracklets have been formed, and a post-processing module named as post association (PA) is employed to associate tracklets to form longer tracks. This PA step is formulated as a Minimum Cost Perfect Matching Problems (MCPM) in general graphs, and the distance of two tracklets is measured as the cosine distance between the average feature of each tracklet. This PA step can be repeated multiple times to progressively form longer tracks. Though this approach is still a sub-optimal solution, it works fairly well when the appearance or motion clues are not reliable in adjacent frames.*

## 1. Introduction

For the the past twenty years, Multi-Objects Tracking (MOT) is mainly the domain that only focuses on tracking single-category objects like people or vehicles in a video. The vast majority of various object categories have been poorly studied. There is an increasing trend to extend MOT algorithms to multiple categories, and to various scenarios, like TAO [3] and OVIS [12] datasets. Besides, as a highly related research direction, video object segmentation (VOS) has already developed similar ideas to segment and track objects of different categories. Towards truly video understanding, it is necessary to perform MOT on different categories and scenes.

To track and segment objects from different categories, we still need to explore many key components (or in an end-to-end manner). During the first TAO challenge [3], **AOA** [4] was the first method to extend the concept of object ReID to truly instance-level appearance representation on the MOT domain and ranked the 1st in the challenge. Following the solution in **AOA**, we extend it for the challenge of RobMOTS, and demonstrate that the anything-appearance model is strong enough for robust MOTS.

## 2. Related Work

Tracking-by-detection (**TBD**) pipeline has been the dominating MOT community. There are three key components in a TBD pipeline: detection, object representation and association. An object detector like Faster RCNN [13] is first used to detect objects from each frame of the sequences. Then, some form of feature embeddings for representing objects are extracted and various association protocols can be applied to get finally connected tracking results. Many works try to improve the object representation ability of objects: DeepSort [20] is a very simple version to associate objects by using Kalman Filter as motion cues and feature embedding model as appearance cues. STRN [22] incorporate those motion and appearance cues into one embedding by a spatial and temporal attention module. Besides, results from the ReID domain [9] can also help to learn better appearance feature for objects. Like AOA [4], our solution also directly borrows the state-of-the-art ReID models to learn to represent any objects.

JDE [18] and FairMOT [24] are newly emerging methods that combine detection and embedding extraction into

a single model, and achieve competitive results with those methods of separated detection and ReID modules.

There are two directions for data association: online and offline association. In online association, a bipartite graph matching problem is formed to match detections and the previous tracklets. Offline methods treat all detections from the whole sequence as nodes of a graph and a Min Cost Network Flow (MCF) problem can be built to solve the matching problem as a whole [23]. Besides, Graph Neural Networks (GNN) have also been applied to solve the matching problem.

Besides TBD, [1] and [27] start another paradigm: tracking-by-regression, which tracks objects by regressing targets based on their positions from previous frames. This paradigm unifies detection and object representation, and also simplifies the association step.

There are also works like [17] which builds an end-to-end neural network to jointly optimize detection and association. In an end-to-end manner, information or gradients from tracking objectives should help on detection step.

## 3. Methods: A simple and robust baseline with existing ingredients

Due to the limited time of the challenge, we choose to simply follow the TBD paradigm. Since we directly re-use the provided strong detection/segmentation results, the feature embedding and data association part are the only missing components to archive a robust solution.

### 3.1. Appearance Modeling: learning appearance embeddings for arbitrary objects

Similar to **AOA**, two appearance models are used in our solutions for general categories except "person" and "car". The baseline model is trained with the State-of-the-Art ReID framework [9] with training data collected from several Single-Object-Tracking datasets. Over 310,000 trajectories are obtained and treated as corresponding instances. This model is denoted as ReID-SOT, which already has a fairly good representation ability. Then ReID-SOT is fine-tuned on the training data of the RobMOTS challenge, making the second model denoted as ReID-Rob-FT. The embeddings of these two models are concatenated together to for an ensemble feature for all objects except "person" or "car".

As for "person" and "car", there are already many dedicated datasets in the ReID community. We trained two models on these two categories respectively with a collection of the public datasets. The person ReID model is trained with almost 23,000 IDs by combining the following datasets: Market1501 [25], DukeMTMC [14], MSMT17 [19], MCT_NLPR [2], CUHK03 [7], EndToEnd [21], RQEN [15], RPIfiled [26], Airport[5]. As for vehicle ReID, the training data is from Veri-wild [8] with over 30,000 IDs.

**Algorithm 1** Algorithm for Cascade Association

**Input:** Detections $\{DET^T\}$, Tracklets $\{TRK^{T-1}\}$
**Output:** Tracklets $\{TRK^T\}$

1: Initialize **Cost Matrix** $M$, **Threshold Embedding** $Tr_f$, **Threshold IoU** $Tr_{IoU}$, **Smooth Control** $alpha = 0.8$, **Life cycle** $lc = 40$
2: **for** $i = 0$ to $len(TRK^{T-1})$ **do**
3:  **for** $j = 0$ to $len(DET^T)$ **do**
4:  Set $M_{ij}$ Calculate **Cosine** distance $c$ of embedding feature between $TRK_i^{T-1}$ and $DET_j^T$
5:  **end for**
6: **end for**
7: Thresholding $M$ with $Tr_f$.
8: Apply Hungarian Algorithm to $M$, get associated detection $DET_a$ and tracklets $TRK_a$
9: Update $TRK_a$ with $DET_a$
10: $DET^T \leftarrow DET^T \setminus DET_a$
11: $TRK^{T-1} \leftarrow TRK^{T-1} \setminus TRK_a$
12: $TRK^T \leftarrow TRK_a$
13: Clean $M$
14: Repeat line 2 to line 8 with IoU as distance and $Tr_{IoU}$, get associated detection $DET_d$ and tracklets $TRK_d$
15: Update $TRK_d$ with $DET_d$
16: $TRK^T \leftarrow TRK^T \cup TRK_d \cup (DET^T \setminus DET_d)$
17: Remove $TRK^{T-1}$ with life cycle $> lc$
18: $TRK^T \leftarrow TRK^T \cup TRK^{T-1}$
19: Smooth Feature by $TRK_i^T = alpha * TRK_i^{T-1} + (1 - alpha) * DET^T$
20: **return** $\{TRK^T\}$

### 3.2. Cascaded Association: from instance-level to tracklet-level matching

The first and core step for cascade association is forming tracklets by matching detections of the current frame with the tracklets already formed in the previous frames. First round of association is based on the concatenated appearance features only. Matching by the appearance with a relatively tight threshold can build up very reliable links with a high precision, and also has the potential to link objects with big displacement caused by object or camera movements. However, there are cases that appearance features would fail to produce reasonable feature similarities, and we introduce a second step of association based on mask IoU. Unlike the provided baseline "STP" which utilizes box IoU, we find that mask IoU yields better results to link the remaining detections and tracklets. The detailed algorithm for the two-step cascaded association is described in Algorithm 1, showing how to perform data association of **Detections** at frame ***T*** ($DET^T$) and **Tracklets** at ***T-1*** ($TRK^{T-1}$) to form **Tracklets** at ***T*** ($TRK^T$).

### 3.3. PA: Tracklet-Level Post Association

Post Association has been first used in [16] as a post-processing technique of improving the tracking performance without much additional cost. After the initial tracklets are formed, each tracklet consists of a list of detection results, each represented by an appearance embedding. Thus, each tracklet can be represented by the average feature of all frames, and the similarity of two tracklets can be obtained as the cosine distance of their averaged feature. Therefore, the Post Association problem is formulated as solving perfect matching on a general graph. In [16] and [4], PA was solved by a greedy algorithm. Here we formulate PA as a Minimum Cost Perfect Matching Problems (MCPM) in general graphs and use [11] as the solver. It should be noted that MCPM here is a sub-optimal solution because it can match between consecutive pairs. If a true track is broken into more than 2 shorter tracklets, MCPM has to be repeated multiple times to get them all connected.

## 4. Experiments

### 4.1. Details of Implementation

All the ReID models are trained with ResNet-50 [6] with IBN-a[10] as backbone and SGD optimizer for 31 epochs. The learning rate decay is performed at epoch 12 and 24, respectively. The dimension of feature embedding for "person" and "car" is **512**. While for other general categories, each model outputs a **256**-dimensional feature so the feature dimension after concatenation is also **512**.

Different values of the threshold $Tr_f$ are adopted for different models/categories, *i.e.* **[0.35, 0.2, 0.5]** for "person", "car" and all other categories respectively. As shown in Algorithm 3.1, the feature of a tracklet is smoothed by **0.8** when a new detection is appended. The life-cycle of a tracklet is setting to **40** (for how long can the tracklet be unobserved, a.k.a. "**MAX_FRAME_SKIP**" in the STP baseline) since robust appearance feature can help with the case when an object re-appears.

### 4.2. Ablation Studies

For simplicity, all ablation results are presented on the validation set, if not specially pointed.

**Comparison of Two Appearance Models** As shown in 1, we reach a better performance when concatenating two ReID models for general categories. Additionally, the one finetuned on the challenge data performs only slightly better compared with the one trained on other public datasets.

**Effectiveness of the Cascaded Association** The top part of Table 2 shows how the performance gets improved by extending the appearance-based association to a cascaded association with mask IoU as the second step and PA as the final step. The bottom part of Table 2 shows that if we switch the order of appearance-based association and mask IoU-based association, the final performance will be degraded even with all other components unchanged. Appearance-First approach is a better choice due to its ability to differentiate objects without any motion or scale assumption, and IoU-based association can serve a good compensation. Besides, feature smoothing is a crucial step as it helps reconnect objects after its occlusion.

| Models | HOTA | AssA | DetA |
|---|---|---|---|
| ReID-Rob-FT | 60.42 | 64.97 | 57.69 |
| ReID-SOT | 60.44 | 65.09 | 57.66 |
| Concatenation | 60.53 | 65.21 | 57.68 |

Table 1. Comparison of two appearance models.

| Steps/Matching Protocols | HOTA | AssA | DetA |
|---|---|---|---|
| Appearance(**APP.**) First | 53.10 | 53.60 | 54.63 |
| + **Mask IoU** | 59.16 | 62.47 | 57.70 |
| + **PA.** | 59.48 | 63.10 | 57.70 |
| + Feature Smoothing(**FS.**) | 60.53 | 65.21 | 57.68 |
| **Mask IoU** First | 55.73 | 56.42 | 57.08 |
| + **App.** + **PA.** + **FS.** | 57.95 | 60.03 | 57.69 |

Table 2. Effectiveness of Cascade Steps.

| Entires | HOTA | AssA | DetA |
|---|---|---|---|
| RobTrack | 61.20 | 64.76 | 59.43 |
| **SBT** | 58.59 | 63.07 | 55.92 |
| SIA | 56.87 | 59.81 | 55.83 |
| MeNToS | 55.52 | 60.80 | 52.38 |
| STP | 54.35 | 55.04 | 55.78 |

Table 3. Top Entries and Baseline (STP) on Test Leaderboard.

### 4.3. Comparison with other entries

A few top entries on the test leaderboard are listed in Table 3. Our proposed approach **SBT** ranked **2nd** among all entries, showing its robustness.

## 5. Conclusion

In this report, we present our solution to the RobMOTS challenge for how to track on 8 benchmarks robustly. As a continuous version of AOA, the final result prove the robustness of the proposed cascade matching methods, the effectiveness of extending single-category ReID to anything appearance modeling on tracking as well as the improved version of post association module. The appearance model should be well studied further in order to archive competitive result to current "person" or "car" ReID.

## References


[1] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[2] Weihua Chen, Lijun Cao, Xiaotang Chen, and Kaiqi Huang. An equalized global graph model-based approach for multi-camera object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(11):2367–2381, 2016. 2

[3] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *European conference on computer vision*, pages 436–454. Springer, 2020. 1

[4] Fei Du, Bo Xu, Jiasheng Tang, Yuqi Zhang, Fan Wang, and Hao Li. 1st place solution to ECCV-TAO-2020: detect and represent any object for tracking. *arXiv preprint arXiv: 2101.08040*, 2021. 1, 3

[5] Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, Richard J Radke, et al. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):523–536, 2018. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[7] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 2

[8] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3235–3243, 2019. 2

[9] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, 2019. 1, 2

[10] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 3

[11] Dilson Lucas Pereira. Algorithms for maximum cardinality matching and minimum cost perfect matching problems in general graphs. https://github.com/dilsonpereira/Minimum-Cost-Perfect-Matching, 2018. 3

[12] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation. *arXiv preprint arXiv:2102.01558*, 2021. 1

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 1

[14] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 2

[15] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[16] Jiasheng Tang, Xiong Xiong, Chenwei Xie, Yanhao Zhang, Pichao Wang, Fan Wang, Fei Du, Liang Han, Yun Zheng, Pan Pan, et al. Min-cost network flow and trajectory fix for multiple objects tracking. In *Conference on Computer Vision and Pattern Recognition Workshop*, volume 2, 2020. 3

[17] Yongxin Wang, Kris Kitani, and Xinshuo Weng. Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. *arXiv:2006.13164*, 2020. 2

[18] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *arXiv preprint arXiv:1909.12605*, 2019. 1

[19] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 2

[20] Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 748–756. IEEE, 2018. 1

[21] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2017. 2

[22] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3988–3998, 2019. 1

[23] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[24] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *arXiv preprint arXiv:2004.01888*, 2020. 1

[25] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 2

[26] Meng Zheng, Srikrishna Karanam, and Richard J Radke. Rpifield: A new dataset for temporally evaluating person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1893–1895, 2018. 2

[27] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 2