Uncertainty-aware Unsupervised Multi-Object Tracking
(Supplementary Material)

Kai Liu[1], Sheng Jin[2], Zhihang Fu[2], Ze Chen[2], Rongxin Jiang[1], Jieping Ye[2]
[1]Zhejiang University, [2]Alibaba DAMO Academy

A. Extended Experiments

A.1. Traking on BDD100K-MOT

To evaluate the generalization of our method, we conduct extended experiments on the BDD100K-MOT [8] dataset. BDD100K-MOT is the largest driving video dataset with 2000 videos in total, needing to track objects of 8 classes. The large camera motion, low frame rate, and fast object moving make it a challenging benchmark in the MOT community. We have recently noticed that ByteTrack [9], the most competitive unsupervised tracker against ours, has achieved the SOTA performance on BDD100K-MOT. For a fair comparison, on each evaluation split, we take the same YOLOX [3] detector and individually perform tracking with ByteTrack and our U2MOT[1]. The results shown in Tab. A1 demonstrate our proposed U2MOT **consistently outperforms** ByteTrack by a large margin, especially on class-averaged terms of mHOTA, mMOTA, and mIDF1, and ID switches. It indicates our U2MOT learned intra-class discriminable feature embedding on all classes.

Specifically, ByteTrack indicates tracking without embedding learning can also bring high performance, while this paper argues the task-specific ReID embedding is still necessary. To handle the multi-class MOT task during inference, ByteTrack adopts an extra pre-trained UniTrack [7] model as the unsupervised ReID embedding extractor [9]. For U2MOT , we add a ReID head, which is trained by our unsupervised framework, on top of the detector. According to Tab. A1, we argue that though ByteTrack has achieved advanced tracking performance with the motion model only (*i.e.*, the Kalman filter), learning task-specific ReID embedding is still necessary for MOT tasks, especially in multi-class situations.

A.2. Discussion on the degradation of MOTA

Tab.1 in our manuscript shows our U2MOT causes a drop of MOTA on MOT17 and MOT20 datasets, compared to ByteTrack [9]. Here we provide more discussions and experiments to figure out this problem.

MOTA is computed as: $\texttt{MOTA} = 1 - \frac{\texttt{FP+FN+IDS}}{\texttt{GT}}$. Compared to ByteTrack [9], for example, though we get a better association-aware ID Switch (IDS) (decreased by 700), the detection-aware False Positive (FP) and False Negative (FN) are worse (increased by 2,300 and 2,500 respectively). Thus the MOTA earns a drop.

A possible reason is that vanilla ByteTrack contains a detector only, while our method adds an extra ReID head. The conflict between the detection head (pulling all targets together) and the ReID head (pushing different targets away) may damage detection performance and ultimately the MOTA [9].

We adopt a naive trick to alleviate this problem: train the detector first, then keep it frozen and train the ReID head. By doing so, the MOTA is increased from 79.7% to 80.2%, which is comparable to 80.3% of ByteTrack. Whereas the HOTA decreased from 64.2% to 63.9%, but still higher than 63.1% of ByteTrack.

In summary, the competition between detection and re-identification tasks deserves be further explorations.

split	Tracker	mHOTA↑	mMOTA↑	mIDF1↑	HOTA↑	MOTA↑	IDF1↑	IDS↓	FPS↑
val	ByteTrack [9]	38.9	33.1	44.2	57.9	61.3	65.8	29083	11.8
	U2MOT (Ours)	**40.7**	**35.5**	**49.1**	**58.7**	**62.9**	**68.9**	**16191**	**19.6**
test	ByteTrack [9]	39.7	28.1	45.9	57.6	61.4	66.1	51979	11.8
	U2MOT (Ours)	**42.2**	**30.7**	**51.2**	**58.3**	**63.0**	**69.2**	**29985**	**19.6**

Table A1: **Evaluation on BDD100K.** On each split, results of ByteTrack and U2MOT are obtained from the SAME detector.

[1] Since ByteTrack's pre-trained model on BDD100K is not released, we re-train the detector according to their codebase and paper details.

B. Uncertainty Derivation

In multi-object tracking, let's consider the association for current object o_i^t with previous $M^{t\text{-}1}$ objects/trajectories ($\{o_1^{t\text{-}1}, o_2^{t\text{-}1}, \cdots, o_{M^{t\text{-}1}}^{t\text{-}1}\}$). After Hungarian algorithm, let $c_{i,j}$ denote the similarity of associated pair ($o_i^t \sim o_j^{t\text{-}1}$), and c_{i,j_2} be the highest similarity against the left objects other than $o_j^{t\text{-}1}$. For a good tracker, $c_{i,j}$ should be close to 1, and c_{i,j_2} is 0. In the whole data space, given $D = \{c_{i,j}^1, c_{i,j}^2, \cdots, c_{i,j}^D \ ; \ c_{i,j_2}^1, c_{i,j_2}^2, \cdots, c_{i,j_2}^D\}$, let $\{p^1, p^2, \cdots, p^D \ ; \ n^1, n^2, \cdots, n^D\}$ be the labels. According to the Multivariate Bernoulli Distribution, the Probability Mass Function is formulated as:

$$P = \prod_{k=1}^{D} \left(c_{i,j}^k\right)^{p^k} \left(1 - c_{i,j}^k\right)^{1-p^k} \left(c_{i,j_2}^k\right)^{n^k} \left(1 - c_{i,j_2}^k\right)^{1-n^k} \tag{A1}$$

The *log-likelihood* is expressed as:

$$\begin{aligned} \log P = \sum_{k=1}^{D} & p^k \log\left(c_{i,j}^k\right) + \left(1 - p^k\right) \log\left(1 - c_{i,j}^k\right) \\ & + n^k \log\left(c_{i,j_2}^k\right) + \left(1 - n^k\right) \log\left(1 - c_{i,j_2}^k\right) \end{aligned} \tag{A2}$$

Considering $p^k = 1$ and $n^k = 0$, it can be simplified as:

$$\log P = \sum_{k=1}^{D} \log\left(c_{i,j}^k\right) + \log\left(1 - c_{i,j_2}^k\right) \tag{A3}$$

To maximize the *log-likelihood*, it is equivalent to minimize the negative *log-likelihood*:

$$\mathcal{L} = \sum_{k=1}^{D} - \log\left(c_{i,j}^k\right) - \log\left(1 - c_{i,j_2}^k\right) \tag{A4}$$

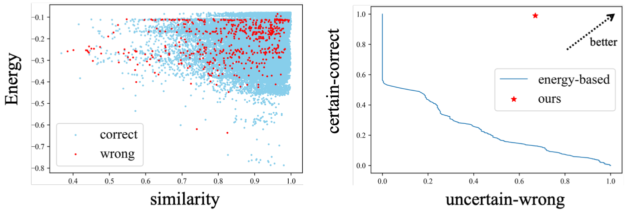
We thus propose to estimate the association risk by:

$$\sigma_{i,j} = -\log\left(c_{i,j}\right) - \log\left(1 - c_{i,j_2}\right) \tag{A5}$$

It is defined as Eq. (4) in our manuscript. According to common sense, when the similarity of associated ($o_i^t \sim o_j^{t\text{-}1}$) is relatively low (*i.e.*, $c_{i,j} < m_1$), or there also exists other similar objects (*i.e.*, $c_{i,j_2} > c_{i,j} - m_2$), the association is uncertain. Furthermore, we have:

$$\begin{aligned} -\log\left(c_{i,j}\right) &> -\log\left(m_1\right) \\ -\log\left(1 - c_{i,j_2}\right) &> -\log\left(1 + m_2 - c_{i,j_2}\right) \end{aligned} \tag{A6}$$

Combining with Eq. (A6), we propose to evaluate the lower bound of Eq. (A5) as:



(a) Energy-based verification.

(b) Verification comparison.

Figure A1: **Statistics of energy-based risk estimation**.

$$\begin{aligned} -\log\left(c_{i,j}\right) - \log\left(1 - c_{i,j_2}\right) &= \sigma_{i,j} \\ > -\log m_1 - \log\left(1 + m_2 - c_{i,j}\right) &\triangleq \gamma_{i,j} \end{aligned} \tag{A7}$$

The margin-based adaptive threshold $\gamma_{i,j}$ is defined as Eq. (5) in the manuscript. When the risk $\sigma_{i,j}$ is higher than the threshold $\gamma_{i,j}$, the assignment ($o_i^t \sim o_j^{t\text{-}1}$) should be considered as **uncertain**. The final quantified uncertainty estimation is formulated by $\delta_{i,j} = \sigma_{i,j} - \gamma_{i,j}$ (Eq. (6)).

C. Uncertainty Manifestation

Other than the proposed estimation of association risk in Eq. (A5), we also explored other manifestations, *e.g.*, the energy model [5]. As the association process can be viewed as a multi-category classification task, the energy score can be expressed as:

$$E_i = -\log \sum_{j=1}^{M^{t\text{-}1}} \exp\left(c_{i,j}\right) \tag{A8}$$

However, as shown in Fig. A1a, the energy score is unable to distinguish the wrong and correct associations by a constant threshold [5]. Quantitatively, energy model failed to concurrently filter the wrong association as uncertain ones and preserve the correct association as certain, as shown in Fig. A1b.

Compared to the energy model, our proposed uncertain metric (risk estimation with adaptive threshold) is more effective and specific for tracking task.

D. Parameter Stability

As our method introduces several hyper-parameters, including m_1/m_2 in Eq.(3) and β/K in Eq.(7), here we provide further validation on the parameter stability on different datasets and base detectors.

In particular, we tested the parameter stability with CenterNet as the base detector on MOT20 validation set. According to Fig. A2, the final performance is insensitive to m_1 and m_2, but relatively sensitive to β and K. It implies our uncertainty measure (involving m_1 and m_2) is stable,

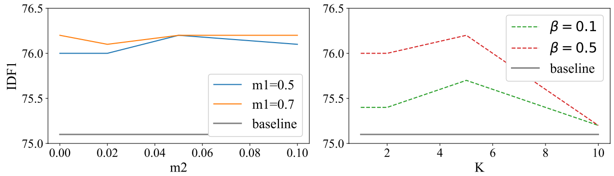


Figure A2: **Evaluation on parameter stability with CenterNet on MOT20.**

though the rectification stage (involving β and K) depends on practical factors like crowd occlusion.

E. Hierarchical Sampling in TGA

To further leverage the uncertainty to improve the inter-frame consistency via the proposed tracklet-guided augmentation, a hierarchical sampling mechanism is developed to select the anchor tracklets and target frames to perform TGA. To illustrate the hierarchical sampling process more clearly, we provide a visualization example in Fig. A3.

F. Quantitative strategy-comparison

To better clarify the advantages of our proposed method, *i.e.*, the Uncertainty-aware Tracklet-Labeling (UTL) mechanism and the Tracklet-Guided Augmentation (TGA) strategy, we provide an extra quantitative comparison with relevant approaches.

Specifically, we have compared UTL with the energy-based [5] uncertainty metric on the MOT17 validation set. Results shown in Tab. A2 indicate our UTL significantly boosts the performance. It is consistent with the visualization in our supplementary materials that such an energy-based metric is inferior to identifying risky associations.

Uncertainty	Augmentation	HOTA↑	IDF1↑	FPS↑
Energy [5]	**TGA**	63.64	74.56	7.4
UTL	FD-GAN [2]	64.85	76.43	2.6
UTL	**TGA**	**64.90**	**76.66**	**7.5**

Table A2: **Module comparison.** 'FPS' is training speed.

Besides, our TGA slightly outperforms such GAN-based adaptive augmentation [2]. And meanwhile, we are 2.5× faster during training. Our tracklet-guided augmentation takes into account both efficacy and efficiency.

G. Cross-Validation for Ablation Studies

MOT17 [6] contains 7 videos in the training set in total. Following the commonly-used ablation protocol [10, 9, 4], we take the first half images of each video for training, and the second half for validation in the manuscript. Under that protocol, the effectiveness of our proposed method is verified. Here we provide more sufficient experimental results to validate our U2MOT through another video-separated ablation protocol.

Specifically, we randomly select 4 videos for training and the other 3 videos for validation. Furthermore, we independently run the video-selection for three times for cross-validation, which are:

- *Training*: MOT17-02, MOT17-09, MOT17-11, MOT17-13. *Validation*: MOT17-04, MOT17-05, MOT17-10.

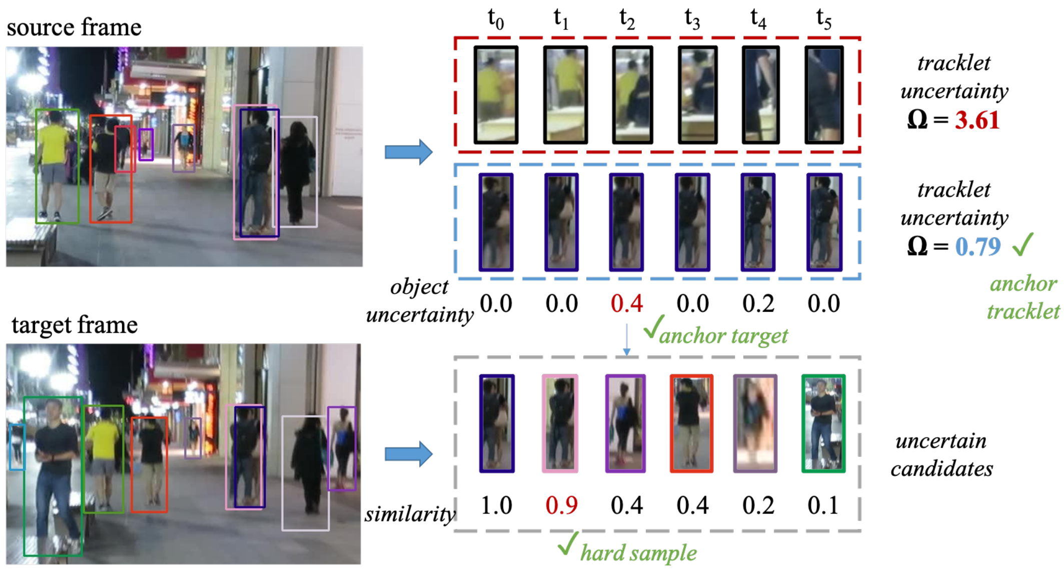


Figure A3: **Visualization of the hierarchical uncertainty-based anchor-sampling mechanism.**

Method	HOTA↑	MOTA↑	IDF1↑	IDS↓
baseline	56.93	65.79	70.31	237
+LTD	57.10	65.82	70.49	236
+UTL	57.26	65.81	70.81	234
+TGA	**57.38**	65.79	**71.02**	**234**
supervised	*57.26*	*65.82*	*70.95*	*234*

Table A3: **Evaluation of the proposed modules.**

m_1	m_1	HOTA↑	MOTA↑	IDF1↑	IDS↓
–	–	57.13	65.78	70.61	237
0.5	0.00	57.30	65.82	70.87	237
0.5	0.02	57.33	65.77	70.99	233
0.5	0.05	**57.38**	65.79	**71.02**	234
0.5	0.10	57.32	65.80	70.98	235
0.7	0.00	57.26	65.80	70.77	237
0.7	0.02	57.36	65.80	70.99	233
0.7	0.05	57.36	65.79	70.98	235
0.7	0.10	57.25	65.83	70.81	235

Table A4: **Ablation on the uncertainty-metric in UTL.** The "–" indicates UTL is not applied.

- *Training*: MOT17-02, MOT17-04, MOT17-05, MOT17-10. *Validation*: MOT17-09, MOT17-11, MOT17-13.
- *Training*: MOT17-05, MOT17-09, MOT17-10, MOT17-13. *Validation*: MOT17-02, MOT17-04, MOT17-11.

Under each protocol, the ablation studies on the proposed components, uncertainty metric, tracklet-guided augmentation are conducted individually. The overall results are shown in Tabs. A3 to A6, where the performance variation are consistent with the ablation protocol in the manuscript. We can draw the same main conclusions: 1) our proposed UTL and TGA is effectice; 2) the uncertainty metric is not relatively sensitive to the hyper-parameters (*i.e.*, the margins); 3) the hierarchical uncertainty-based sampling mechanism further improves the augmentation's quality; and 4) UTL is a generalized plug-and-play module that can be integrated into existing methods and consistently improve the tracking performance. The effectiveness of our proposed method is further demonstrated.

H. ReID Head Implementation

For reproducibility, we show a simple implementation of the ReID head in Fig. A4, with which all the results in the manuscript are obtained. Following prior arts [10, 7], we use the multi-scale features to boost the appearance embedding, which is sampled at object's center in the ReID feature map. Such a lightweight ReID head is able to learn objects' consistent feature over the time period.

TGA-src	TGA-tgt	HOTA↑	MOTA↑	IDF1↑	IDS↓
–	–	57.26	65.81	70.81	234
random	random	57.25	65.81	70.88	235
uncertain	random	57.30	65.77	70.91	233
random	uncertain	57.29	65.80	70.95	233
uncertain	uncertain	**57.38**	65.79	**71.02**	234

Table A5: **Ablation on the anchor-selection mechanism in TGA.** The "–" indicates TGA is not applied.

Tracker	HOTA↑	MOTA↑	IDF1↑	IDS↓
baseline	57.38	65.79	71.02	234
ByteTrack	56.89	65.68	70.53	241
+UTL	**58.06**	**65.89**	**72.47**	**237**
FairMOT	55.67	64.62	68.80	672
+UTL	**57.17**	**65.15**	**70.61**	**545**
DeepSORT	52.08	61.92	63.15	604
+UTL	**53.26**	**62.45**	**63.94**	**480**
MOTDT	53.84	63.40	66.17	713
+UTL	**55.01**	**63.85**	**68.04**	**489**

Table A6: **Inference boosting.** Results are obtained by different association strategies with the SAME model.

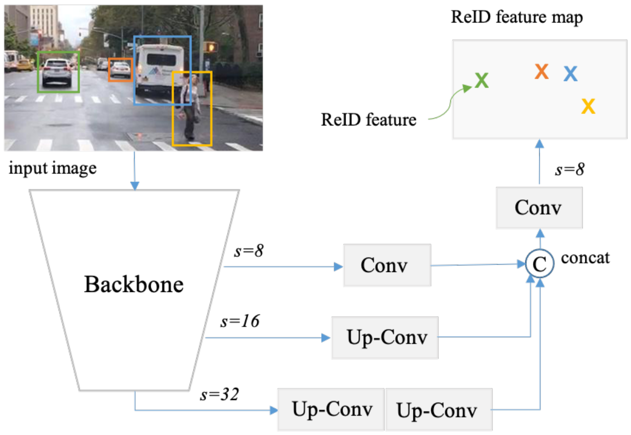


Figure A4: **A simple implementation of the ReID head.** Embedding vector is sampled at object's center in the ReID feature map.

I. Uncertainty Visualization

To further demonstrate the effectiveness of our proposed association-level uncertainty metric, more visualization examples are provided. According to Fig. A5, the uncertainty metric is capable to detect the wrong associations under the following circumstances:

- **Severe object occlusion**. As show in Fig. A5a, when an object is almost entirely occluded by another object, the two objects have similar appearance embedding as well as similar motion information, which make the association ambiguous.
- **Similar object appearance**. As show in Fig. A5b, when two objects are not occluded, the similar appearance still leads to ID switches. Especially, the small objects or blurred appearance will increase the embedding ambiguity.
- **Irregular camera motion**. As show in Fig. A5c, when the camera is irregularly or swiftly moved, the IoU information is no more reliable, especially with ambiguous appearance embeddings.

At the same time, correct associations are preserved as 'certain' ones, which demonstrates the robustness of our uncertain metric.

J. Tracking Visualization

As shown in Fig. A6 and Fig. A7, more visualization examples on MOT17 [6], MOT20 [1], VisDrone-MOT [11], and BDD100K-MOT [8] datasets are provided to further demonstrate the effectiveness of our proposed unsupervised MOT method.

References

[1] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 5

[2] Yixiao Ge. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*, 2018. 3

[3] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1

[4] Yu-Lei Li. Unsupervised embedding and association network for multi-object tracking. In *Proceedings of the 31th International Joint Conference on Artificial Intelligence, (IJCAI-22)*, 2022. 3

[5] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *NeurIPS*, 33:21464–21475, 2020. 2, 3

[6] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 3, 5

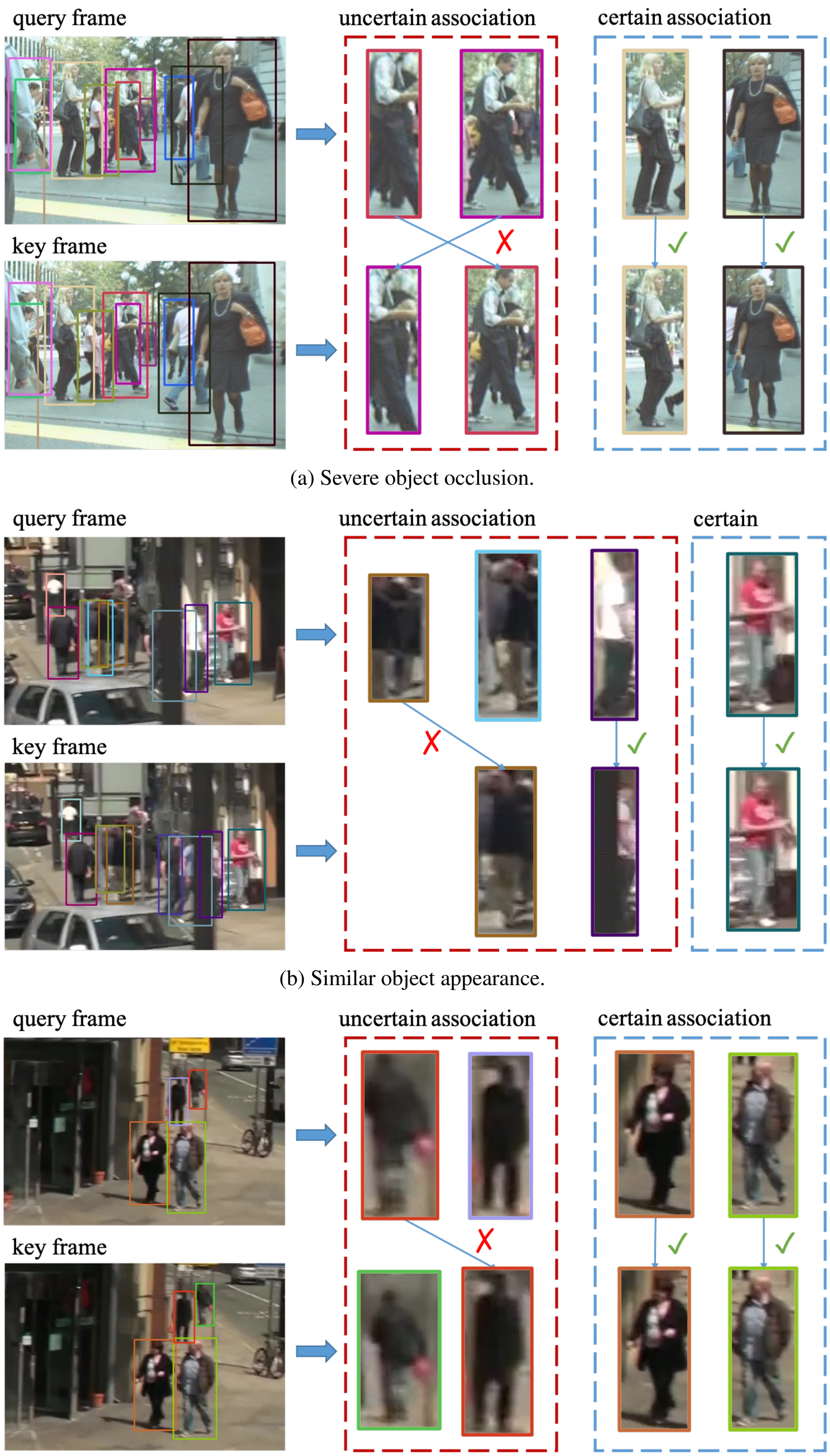
[7] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *NeurIPS*, 34:726–738, 2021. 1, 4

[8] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2020. 1, 5

[9] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 3

[10] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087, 2021. 3, 4

[11] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 5

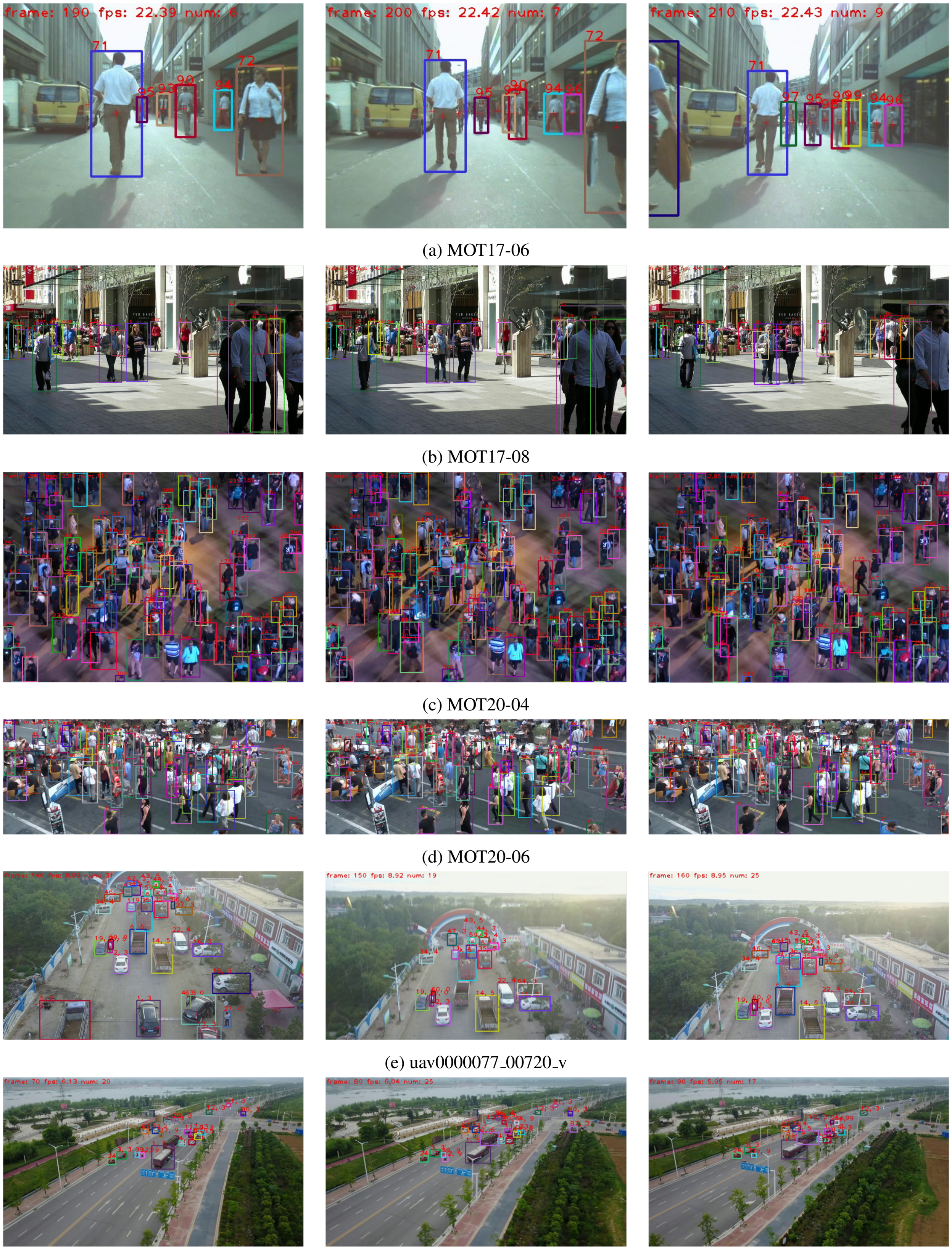


(a) Severe object occlusion.

(b) Similar object appearance.

(c) Irregular camera motion.

Figure A5: **Typical visualizations for uncertain associations.** In each sub-figure, the query (current) frame and its objects are placed at the first row, and the second row display the key (previous) frame and its objects. The object association results are presented as arrows.


frame: 190 fps: 22.39 num: 6
71
95
93
90
94
72
frame: 200 fps: 22.42 num: 7
71
95
90
94
96
72
frame: 210 fps: 22.43 num: 9
71
97
95
98
90
99
94
96
frame: 150 fps: 8.92 num: 19
frame: 160 fps: 8.95 num: 25
frame: 70 fps: 6.13 num: 20
frame: 80 fps: 6.04 num: 25
frame: 90 fps: 5.95 num: 17

(a) MOT17-06

(b) MOT17-08

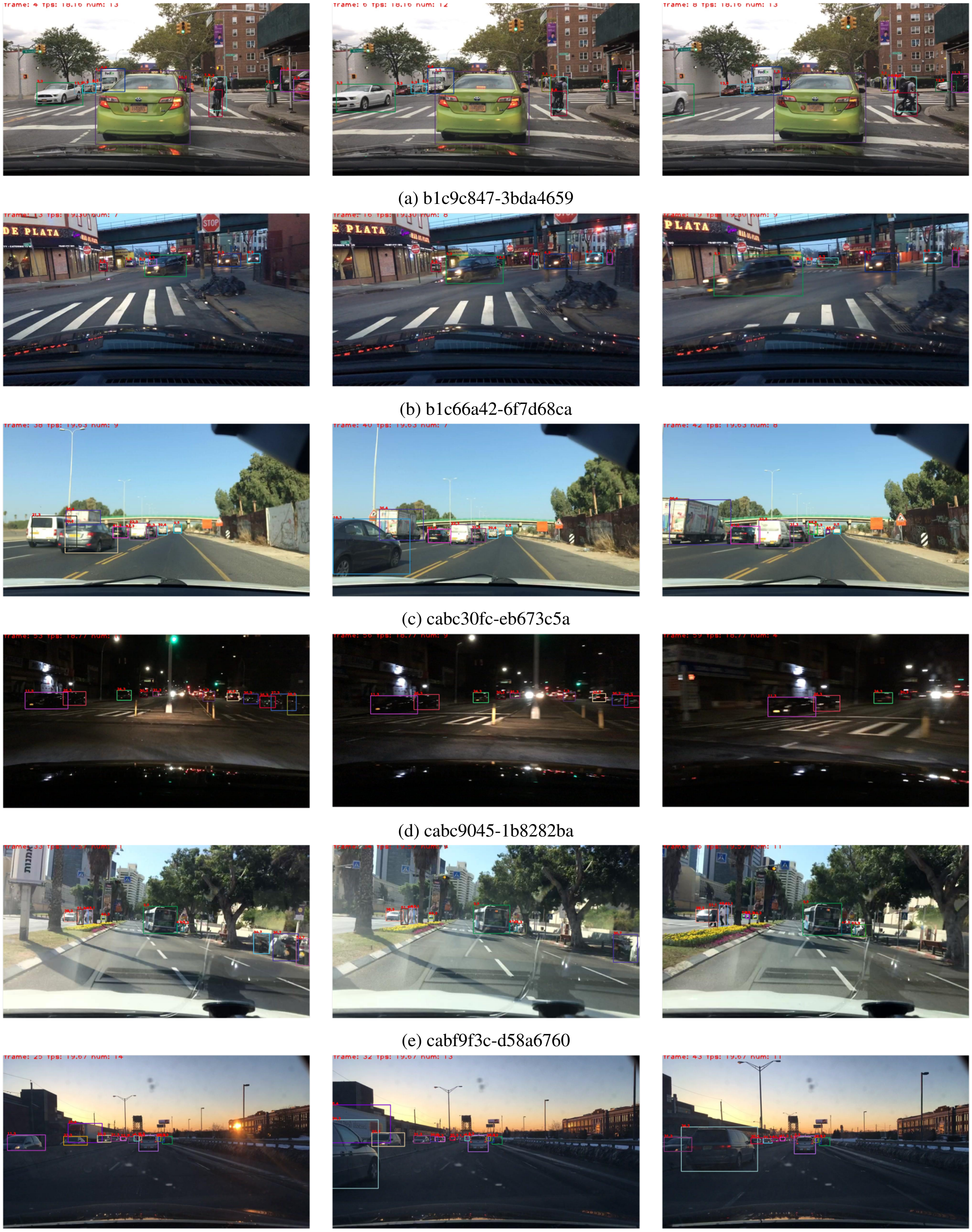
(c) MOT20-04

(d) MOT20-06

(e) uav0000077_00720_v

(f) uav0000201_00000_v

Figure A6: **Typical visualizations of our unsupervised multi-object tracking results.** In MOT-17 ((**a-b**)) and MOT-20 ((**c-d**)), different colors represent different *identities*. In VisDrone-MOT ((**e-f**)), both *identities* and *categories* are presented.


frame: 4 fps: 18.16 num: 13
frame: 6 fps: 18.16 num: 12
frame: 8 fps: 18.16 num: 13
frame: 13 fps: 19.30 num: 7
frame: 16 fps: 19.30 num: 8
frame: 19 fps: 19.30 num: 9
frame: 38 fps: 19.63 num: 9
frame: 40 fps: 19.63 num: 7
frame: 42 fps: 19.63 num: 8
frame: 53 fps: 18.77
frame: 56 fps: 18.77 num: 9
frame: 59 fps: 18.77 num: 4
frame: 25 fps: 19.67 num: 14
frame: 32 fps: 19.67 num: 13
frame: 43 fps: 19.67 num: 11

(a) b1c9c847-3bda4659

(b) b1c66a42-6f7d68ca

(c) cabc30fc-eb673c5a

(d) cabc9045-1b8282ba

(e) cabf9f3c-d58a6760

(f) cabf9f3c-d58a6760

Figure A7: **Typical visualizations of our unsupervised multi-object tracking results on BDD100K-MOT dataset.** Our U2MOT can handle the challenges in autonomous driving scenes, such as various scenarios and diverse motions.