Uncertainty-aware Unsupervised Multi-Object Tracking

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

Abstract

Without manually annotated identities, unsupervised multi-object trackers are inferior to learning reliable feature embeddings. It causes the similarity-based inter-frame association stage also be error-prone, where an ***uncertainty*** *problem arises. The frame-by-frame accumulated uncertainty prevents trackers from learning the consistent feature embedding against time variation. To avoid this uncertainty problem, recent self-supervised techniques are adopted, whereas they failed to capture temporal relations. The inter-frame uncertainty still exists. In fact, this paper argues that though the uncertainty problem is inevitable, it is possible to leverage the uncertainty itself to improve the learned consistency in turn. Specifically, an uncertainty-based metric is developed to verify and rectify the risky associations. The resulting accurate pseudo-tracklets boost learning the feature consistency. And accurate tracklets can incorporate temporal information into spatial transformation. This paper proposes a tracklet-guided augmentation strategy to simulate the tracklet's motion, which adopts a hierarchical uncertainty-based sampling mechanism for hard sample mining. The ultimate unsupervised MOT framework, namely U2MOT, is proven effective on MOT-Challenges and VisDrone-MOT benchmark. U2MOT achieves a SOTA performance among the published supervised and unsupervised trackers.*

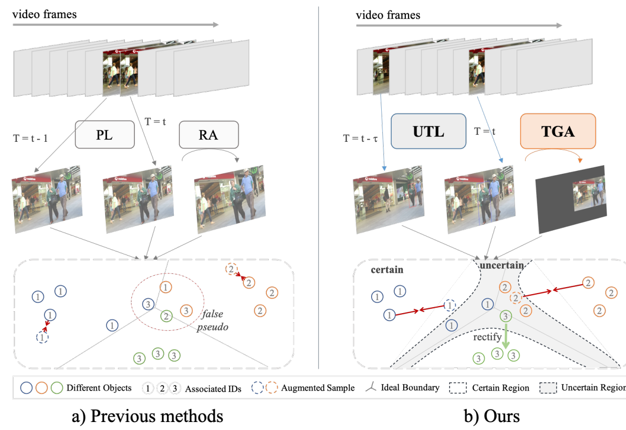


Figure 1: **Method Comparison.** a) Previous methods utilize pseudo-labeled (PL) adjacent frames or randomly augmented (RA) samples. b) Our method adopts uncertainty-aware tracklet-labeling (UTL) to maintain consistent pseudo-tracklets, and exploits tracklet-guided augmentation (TGA) to improve the embedding consistency

1. Introduction

Multi-object tracking (MOT) [31, 4, 46] has been widely deployed in real-world applications, including surveillance analysis [32, 56], autonomous driving [13, 38], intelligent robots [34, 3], *etc*. The goal of MOT task is to detect all target objects and simultaneously keep their respective feature embeddings **consistent**, regardless of the change of their shapes and angles over a period of time [46, 55]. However, the core issue of unsupervised MOT task is lacking the annotated ID-supervision to confirm the consistency of a certain target, especially when its shape and angle are varied over time [19, 43, 24, 22, 37]. When training an unsupervised tracker, since the learned feature embedding is unreliable, the similarity-based association stage is error-prone. Propagating pseudo identities frame-by-frame leads to uncertainty in the resulting pseudo-tracklets, which accumulates in per-frame associations. This prevents trackers from learning a consistent feature embedding [43, 24].

To avoid this problem, self-supervised techniques [24, 37, 52] are utilized to generate augmented samples with perfectly-accurate identities. However, these commonly-used methods merely take a single frame for augmentation, while the inter-frame temporal relation is totally ignored. It usually leads to sub-optimal performance [55]. The uncertainty problem seems inevitable but remains under-explored. In fact, we argue that the uncertainty can be leveraged to maintain consistency in turn, as shown in Fig. 1.

First, uncertainty can guide the construction of pseudo-tracklets. When the similarity-based inter-frame object association is inaccurate, we propose to introduce a quan-

*This work was done when Kai Liu worked as a research intern at Alibaba DAMO Academy. Email: kail@zju.edu.cn.
†Corresponding author. Email: zhihang.fzh@alibaba-inc.com.

tified uncertainty measure to find out the possibly wrong associations and further re-associate them. Specifically, we find that the association mismatching is accompanied by two phenomenons, including a small similarity margin (similar appearance *etc*.) and low confidence (object occlusion *etc*.). Inspired by these findings, an **association-uncertainty** metric is proposed to filter the uncertain candidate set, which is further rectified using the tracklet appearance and motion clues. The proposed mechanism, termed **Uncertainty-aware Tracklet-Labeling (UTL)**, generates highly-accurate pseudo-tracklets to learn the embedding consistency. The proposed UTL has two features: (1) it can directly boost the tracking performance during inference as well. (2) it is complementary to existing methods and can be incorporated with consistent performance.

Second, uncertainty can guide the hard sample augmentation. The trustworthy pseudo-tracklets can be exploited to incorporate temporal information into sample augmentation, thereby overcoming the key limitation of current augmentation-based methods. To this end, we develop a **Tracklet-Guided Augmentation (TGA)** strategy to simulate the real motion of pseudo-tracklets. Specifically, TGA generates augmentation samples aligned to the highly-uncertain objects in the pseudo-tracklet for hard example mining. Because a high association-uncertainty basically indicates the presence of challenging negative examples. To achieve this goal, a hierarchical uncertainty-based sampling mechanism is developed to ensure a trustworthy pseudo-tracklet and hard sample augmentation.

The ultimate unsupervised MOT framework, namely U2MOT, is proved effective on several public benchmarks (*i.e*., MOT17 [31], MOT20 [7], and the challenging VisDrone-MOT [57]). The experiments show that U2MOT significantly outperforms previous unsupervised methods (*e.g*., 62.7% $v.s.$ 58.6% of HOTA on MOT20), and achieves SOTA (*e.g*., 64.2% HOTA on MOT17) among existing unsupervised and supervised trackers. Extensive ablation studies demonstrate the effectiveness of leveraging uncertainty in improving the consistency in turn.

Contributions of this paper are summarized as follows:

1) We are the first to leverage uncertainty in unsupervised multi-object tracking, where an association-level uncertain metric is introduced to verify the pseudo-tracklets, and a hierarchical uncertainty-based sampling mechanism is developed for hard sample generation.

2) We propose a novel unsupervised U2MOT framework, where UTL is developed to guarantee the intra-tracklet consistency and TGA is adopted to learn the consistent feature embedding.

3) We achieve a SOTA tracking performance among existing methods, and demonstrate the generalized application prospects for the uncertainty metric.

2. Related Work

Pseudo-label-based Unsupervised MOT. Existing unsupervised methods generate pseudo-identities in three main ways, including motion-based, cluster-based, and similarity-based methods. In terms of motion-based methods, SimUMOT [19] adopts SORT [4] to generate the pseudo-tracklets, which is used to guide the training of re-identification networks. Very recently, UEANet [22] uses ByteTrack [54] to improve the quality of pseudo labels, where ByteTrack excavated the values of low-confident detection boxes. However, long-term dependency within pseudo-tracklets is hard to guarantee, and the spatial information is not reliable in irregular camera motions. Cluster-based methods [10, 23, 37] try to iteratively cluster the objects in the whole video to get pseudo-identities. These methods usually lead to sub-optimal performance. A possible reason is that the temporary association within the tracklet is totally ignored. The similarity-based methods, like Cycas [43] and OUTrack [24], utilize the cycle-consistency [42] of object similarities between adjacent frames. As time interval extends, the noise of pseudo-label becomes an inconvenient truth. Different from existing methods, our U2MOT designs an uncertainty-based refinement mechanism to obtain accurate associations. Long-term consistency is preserved through identity propagation.

Uncertainty Estimation. In recent years, uncertainty estimation has been widely explored in classification calibration (*e.g*., detecting misclassified or out-of-distribution samples) from three main aspects. Some researchers adopt deterministic networks [29, 35, 9] or ensembles [21, 45] to explicitly represent the uncertainty. Others adopt the widely-used softmax probability distribution [17] to evaluate the credibility according to the classification confidence. Very recently, the energy model [26, 39] emerges as the widely-exploited metric in the uncertainty estimation, which is theoretically aligned with the probability density of the inputs. However, for multi-object tracking, object occlusion and similar appearance always lead to mismatching. Thus, the uncertain estimation is worth exploring. In this paper, we design an uncertain metric specially for tracklets-based tasks, which is proved effective.

Augmentation Strategy. Adaptive augmentation strategies have been extensively studied in image classification [11, 27], object detection [40, 14], and representation learning [1, 53]. However, random perspective transformation still dominates in unsupervised multi-object tracking [55, 37]. Other researchers present GAN-based augmentation strategies [18, 52] for person re-identification. However, these methods fail to generate realistic object tracklets in MOT situations. This work integrates the tracklets property into augmentation and focuses on negative hard sample generations, which makes our augmentation strategy task-specific and effective.

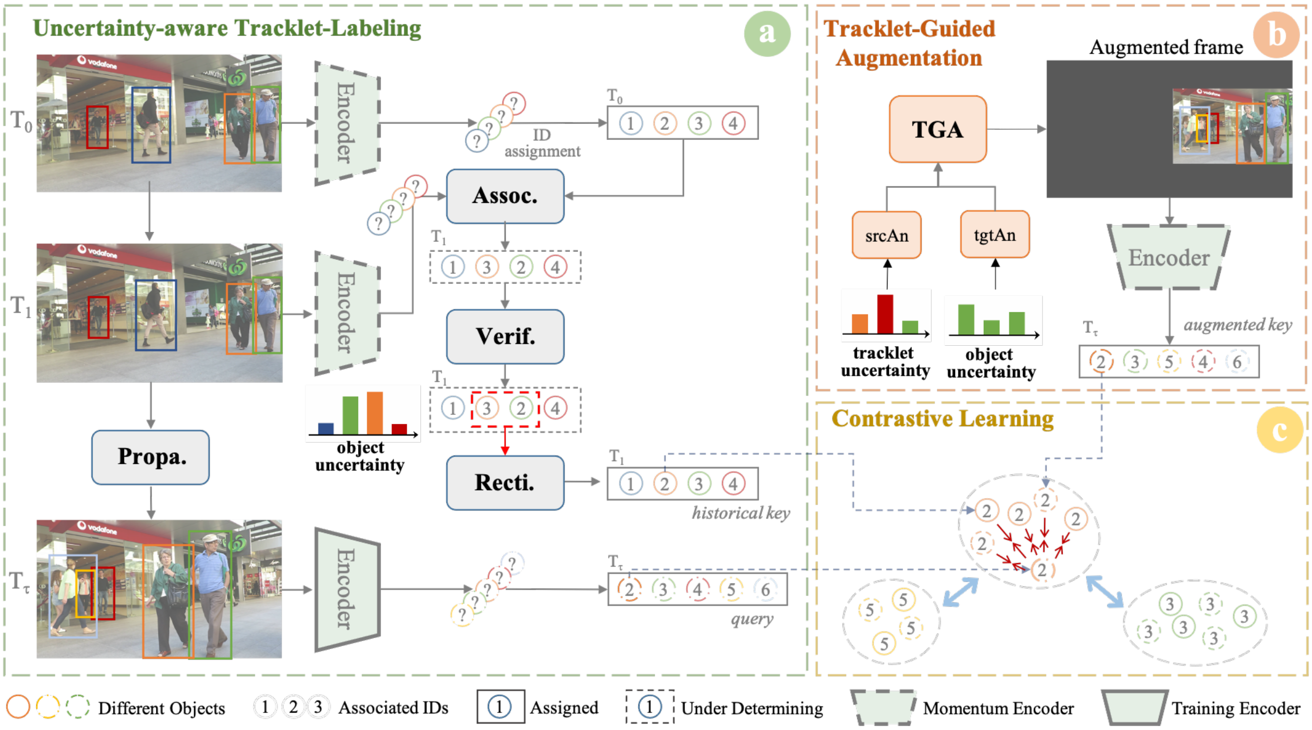


Figure 2: **Framework of U2MOT** (better viewed in color). **a)** We propose an uncertainty metric to verify and rectify the association process. Accurate pseudo-tracklets are propagated frame-by-frame. **b)** Then an anchor pair is selected based on tracklet-level and object-level uncertainties. Tracklet's motion information is used to guide the augmentation. **c)** Contrastive learning is adopted to train the tracker, which pulls the objects within the tracklet together and pushes different tracklets apart.

3. Methodology

3.1. Overview

As shown in Fig. 2, the proposed unsupervised MOT framework is trained with the widely-used contrastive learning technique [6, 15]. Specifically, for multi-object tracking, objects within the tracklet ($\boldsymbol{k}_+$) should be pulled together and different tracklets ($\boldsymbol{k}_-$) should be separated. It can be mathematically formulated as:

$$\mathcal{L}_{cl}(\boldsymbol{q};\boldsymbol{k}_+;\boldsymbol{k}_-) = -\log \frac{\exp(\boldsymbol{q}\cdot\boldsymbol{k}_+/\epsilon)}{\sum_i \exp(\boldsymbol{q}\cdot\boldsymbol{k}_i/\epsilon)} \tag{1}$$

where $\mathcal{L}_{cl}$ denotes the InfoNCE [33] loss function, k_i means a (positive or arbitrarily negative) key sample, and $\epsilon = 0.07$ is the temperature [48]. Following the unsupervised tracking fashion [24, 37], the positive and negative keys mainly come from two sources within a video, *i.e.* pseudo-labeled historical frames and self-augmented frames.

However, two issues occur: (1) the uncertainty reduces the accuracy of pseudo-tracklets and (2) the randomly augmented samples fail to learn the inter-frame consistency. We argue the above issues are not independent. By leveraging the uncertainty in turn, the accurate pseudo-tracklets can guide the qualified positive and negative augmentations.

To address these two issues, we propose an uncertainty-aware pseudo-tracklet labeling strategy in Sec. 3.2, which integrates a verification-and-rectification mechanism into the tracklet generation. Then we propose a tracklet-guided augmentation strategy in Sec. 3.3, bringing temporary information into spatial augmentation. The augmented samples simulate the objects' motion. A hierarchical uncertainty-based sampling strategy is proposed for hard sample mining. More details are described in the following section.

3.2. Uncertainty-aware Tracklet-Labeling

Accurate pseudo tracklet is critical in learning feature consistency. However, without manual annotation, the aggravated uncertainty makes the tracklet-labeling a huge challenge due to various interference factors, including similar appearance among objects, frequent object cross and occlusions, *etc*. In fact, the uncertainty can also be leveraged to improve the pseudo-accuracy in turn. In this section, we propose an **U**ncertainty-aware **T**racklet-**L**abeling (**UTL**) strategy for better pseudo-tracklets.

Given an input video sequence $V = \{I^1, I^2, \cdots, I^N\}$, each frame I^t is annotated with the bounding boxes $B^t = \{b_1^t, b_2^t, \cdots, b_{M^t}^t\}$ of M^t objects in t_{th} frame, where $b_i^t =$

$(cx_i^t, cy_i^t, w_i^t, h_i^t)$ is the center coordinate and shape of the i_{th} object o_i^t. As shown in Fig. 2, U2MOT generates accurate pseudo-tracklets in four main steps:

1) **Association**. For a certain object o_i^t in frame I^t, the ℓ_2-normalized representation $\boldsymbol{f}_i^t$ can be expressed as $\boldsymbol{f}_i^t = \phi(I^t, b_i^t)$, where the embedding encoder is denoted as ϕ.

To associate the objects in frame I^t with the objects or trajectories in previous $I^{t\text{-}1}$, a similarity matrix is constructed with their appearance embeddings:

$$\boldsymbol{C} \in \mathbb{R}^{M^t \times M^{t\text{-}1}}, \; c_{i,j} = \boldsymbol{f}_i^t \cdot \boldsymbol{f}_j^{t\text{-}1} \tag{2}$$

where $c_{i,j}$ represents the cosine similarity between the i_{th} object in frame I^t and the j_{th} object (or trajectory) in frame $I^{t\text{-}1}$. Then the Hungarian algorithm [20] is adopted to generate the identity association results.

2) **Verification**. However, the appearance representations are sometimes unreliable, especially in the unsupervised scenario. To solve this issue, an uncertainty metric is proposed to evaluate the association after the first stage.

Object association can be viewed as a multi-category classification problem. And confidence-score has been proved efficient and effective in detecting mis-classified examples [17]. Inspired by this, we propose to detect the mis-associated objects through the similarity scores.

Given an object o_i^t associated with $o_j^{t\text{-}1}$ in the previous frame based on Eq. (2), the association ($o_i^t \sim o_j^{t\text{-}1}$) is unconvincing in two cases: 1) the assigned similarity $c_{i,j}$ is relatively low (*e.g.*, partial occlusion or motion blur) and 2) there are other objects whose similarities are close to the assigned $c_{i,j}$ (*e.g.*, similar appearance or indistinguishable embedding). It can be formulated as:

$$c_{i,j} < m_1; \quad c_{i,j_2} > c_{i,j} - m_2 \tag{3}$$

where m_1, m_2 are constant margins. For simplicity, only the second-highest similarity with others (c_{i,j_2}) is considered. In an ideal association, $c_{i,j}$ should be close to 1 and c_{i,j_2} close to 0. We thus estimate the association risk as:

$$\sigma_{i,j} = -\log c_{i,j} - \log\left(1 - c_{i,j_2}\right) \tag{4}$$

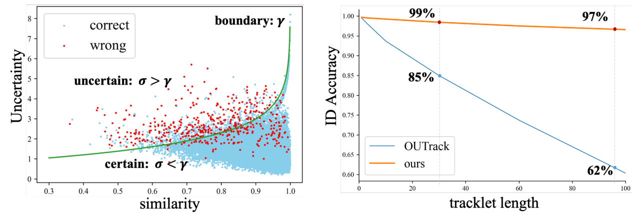
Detailed derivation is shown in Appendix B. Combining with Eq. (3) and Eq. (4) , an adaptive threshold is proposed:

$$\gamma_{i,j} = -\log m_1 - \log\left(1 + m_2 - c_{i,j}\right) \tag{5}$$

As shown in Fig. 3a, 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 reconsidered. The **association uncertainty** is quantified as:

$$\delta_{i,j} = \sigma_{i,j} - \gamma_{i,j} \tag{6}$$

The results are not sensitive to the exact margins. We set $m_1 = 0.5$ and $m_2 = 0.05$ for slightly better performance.



(a) Association verification. (b) Pseudo Accuracy.

Figure 3: **Statistics of pseudo-identities on MOT17-train set**. (a) 67% wrong associations are divided into the 'uncertain' area, and 99% correct associations are preserved in 'certain' area. (b) The pseudo-accuracy of previous methods dramatically drops as tracklet length increases. By contrast, our U2MOT consistently maintains a high accuracy .

The uncertain pairs after the verification stage and unmatched objects after the association stage are gathered as uncertain candidates for the rectification stage. We have provided several visualization examples to verify the certain/uncertain associations in Appendix I.

3) **Rectification**. The rectification stage is performed among the uncertain candidate. The similarities between the two adjacent frames are no longer convincing. More information should be taken into account, including motion estimation and appearance variation within a tracklet.

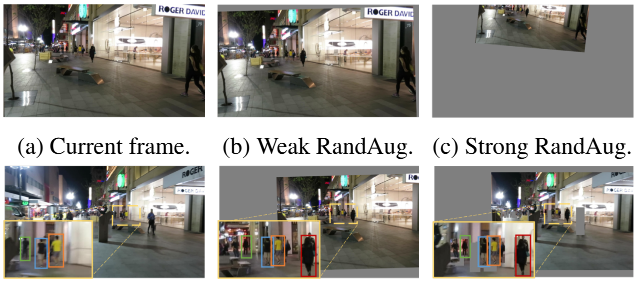
For the uncertain candidates, U2MOT constructs another similarity matrix for the secondary rectification. First, the motion constraints should be relaxed, so the association shares overlap higher than β are preserved. Second, the appearance should not vary extremely fast, so we adopt the averaged similarity between object o_i^t and tracklet $trk_j = \{o_j^{t\text{-}K}, \cdots, o_j^{t\text{-}1}\}$ within previous K frames. In this stage, we solve the sub-problem of global identity assignments, which can be formulated as:

$$\boldsymbol{C}' \in \mathbb{R}^{M^{t\prime} \times M^{t\text{-}1\prime}}$$
$$c'_{i,j} = \left(\frac{1}{K} \sum_{\hat{t}=t\text{-}K}^{t\text{-}1} \boldsymbol{f}_i^t \cdot \boldsymbol{f}_j^{\hat{t}} \right) \times \mathbb{I}\left(\text{IoU}\left(b_i^t, b_j^{t\text{-}1}\right) > \beta\right) \tag{7}$$

where $\mathbb{I}(*)$ is the indicator function. Then the match set is updated based on the Hungarian algorithm.

Remark. Our core contribution is the uncertainty-based verification mechanism, rather than the specific rectification, which shall be adjusted in practice. Empirically we set $\beta = 0.1$ and $K = 5$.

4) **Propagation**. The pseudo-tracklets are propagated frame-by-frame. As shown in Fig. 3b, our strategy brings consistently accurate pseudo-identities, *e.g.*, reaching 97% accuracy across 100 frames. The long-term intra-tracklet consistency is successfully maintained.



(a) Current frame. (b) Weak RandAug. (c) Strong RandAug.

(d) Historical frame. (e) TGA: type-a. (f) TGA: type-b.

Figure 4: **Comparison of augmentations.** Random augmentations (**(b)** and **(c)**) fail to simulate the tracklet movements between frames (**(a)** and **(d)**), while TGA can. Two types of TGA are displayed: current objects + historical position + current **(e)** or historical **(f)** background.

It is worth mentioning that the verification and rectification stages can be naturally applied to the inference process to boost the performance, which does not conflict with existing association methods.

3.3. Tracklet-Guided Augmentation

Accurate pseudo-tracklets can guide the sample augmentation in the unsupervised MOT framework. To learn the inter-frame consistency [6, 55], good training samples should be diverse and temporal-aware. However, as illustrated in Fig. 4, existing methods usually treat augmentation and multi-object tracking as two isolated tasks, leading to ineffective augmentations. Instead, this paper utilizes the tracklet's spatial displacements to guide the augmentation process. Based on a properly selected anchor pair, the proposed strategy makes the augmented frames aligned to the historical frames, simulating realistic tracklet movements. The proposed method concurrently focuses on the hard negative samples. Details of the **T**racklet-**G**uided **A**ugmentation (TGA) are given below.

Given a current frame I^t with M^t objects, a source-anchor object o_a^t is selected, whose bounding box is denoted as $b_a^t = (cx_a^t, cy_a^t, w_a^t, h_a^t)$. Then, we choose a target-anchor $o_a^{t\text{-}\tau}$ in $(t\text{-}\tau)_{th}$ historical frame from the pseudo-tracklet $trk_a = \{o_a^{t_0}, o_a^{t_1}, \cdots, o_a^t\}$. Finally, to augment the current I^t to align with historical $I^{t\text{-}\tau}$, a tracklet-guided affine transformation can be expressed as:

$$\begin{bmatrix} x^{t\text{-}\tau} \\ y^{t\text{-}\tau} \\ 1 \end{bmatrix} = \boldsymbol{M}_t^{t\text{-}\tau} \begin{bmatrix} x^t \\ y^t \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x^t \\ y^t \\ 1 \end{bmatrix} \quad (8)$$

where x^*, y^* are spatial coordinates, and $\boldsymbol{M}_t^{t\text{-}\tau}$ can be solved by direct linear transform (DLT) algorithm [8]. Then an augmented frame $\tilde{I}^t$ is generated based on the tracklet-guided affine transformation with perspective jitter, which can be expressed as $\tilde{I}^t = \mathcal{T}(I^t, M_t^{t\text{-}\tau})$.

Intuitively, a proper anchor-selection is vitally important for our augmentation strategy.

First, the identity accuracy of anchor pair $(o_a^t \sim o_a^{t\text{-}\tau})$ is important. In other words, the consistency of anchor tracklet trk_a should be guaranteed. We thus design a tracklet-level uncertain metric based on the propagated association-level uncertainty defined in Eq. (6), which is formulated as:

$$\Omega_i = \frac{1}{n} \sum_{s=t_0}^{t} \exp(\delta_i^s) \quad (9)$$

where Ω_i denotes the uncertainty of tracklet trk_i, and n is the tracklet length. An uncertainty-based sampling strategy is designed to select the source anchor o_a^t (along with the anchor trk_a) from the M^t objects in frame I^t by:

$$p(a = i \mid t) = \frac{\exp(-\Omega_i)}{\sum_{\hat{i}=1}^{M^t} \exp(-\Omega_{\hat{i}})} \quad (10)$$

where $p(a = i \mid t)$ represents the probability to choose the i_{th} tracklet trk_i as the anchor trk_a. The uncertain tracklet with high Ω is less likely to be selected, avoiding dramatic augmentations from erroneous pseudo-tracklets.

Second, hard negative samples matter in discriminability learning. We tend to choose an indistinguishable (or, highly uncertain) target anchor $o_a^{t\text{-}\tau}$ along the tracklet trk_i. The selection probability can be formulated as:

$$p(\pi = t\text{-}\tau \mid a) = \frac{\exp(\delta_a^{t\text{-}\tau})}{\sum_{\hat{\tau}=t_0}^{t-1} \exp(\delta_a^{t-\hat{\tau}})} \quad (11)$$

Compared to conventional random transformation, our tracklet-guided augmentation is well-directed and tracking-related. A visualization example is displayed in the Appendix E to illustrate the hierarchical sampling process.

Together with accurate pseudo-tracklets, the inter-frame consistency is successfully improved, as shown in Fig. 5.

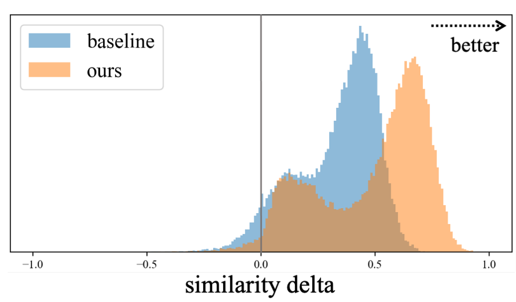


Figure 5: **Inter-frame consistency visualization.** During the associations, we statistic the similarity delta (Δ) between ground-truth association (c^+) and other objects with the largest similarity (c^-). A positive delta ($\Delta = c^+ - c^- > 0$) means a good tracker, the larger the better.

4. Experiment

4.1. Datasets and Evaluation Metrics

Datasets. Experiments are performed on three popular benchmarks: MOT17 [31], MOT20 [7], and the challenging VisDrone-MOT [57]. MOT17 contains 14 videos captured from diverse viewpoints and weather conditions, while MOT20 consists of 8 videos on crowded scenes with heavier occlusion. Both of them are evaluated with the "private detection" protocol. VisDrone-MOT is captured by UAVs in various scenes, which comprises 79 sequences with 10 object categories. Only five categories (*i.e.*, car, bus, truck, pedestrian, and van) are considered during evaluation [25]. Multi-class tracking with irregular camera motion makes VisDrone-MOT a challenging benchmark.

Metrics. Following the previous MOT methods [22, 55, 54], the HOTA [28] and CLEAR [2] metrics are adopted to evaluate the trackers. Specifically, CLEAR mainly includes multiple object tracking accuracy (MOTA), ID F1 score (IDF1), and identity switches (IDS).

4.2. Implementation Details

To show the efficacy of our unsupervised MOT framework, we implement U2MOT on YOLOX [12] with a ReID head integrated (see Appendix H). Specifically, the detection branch is trained in the same way as ByteTrack [54], while the new ReID head is learned with our U2MOT. The model is trained with SGD optimizer and the initial learning rate of 1×10^{-3} with cosine annealing schedule.

For MOT17 and MOT20, the model is trained with the same setting as ByteTrack. Taking MOT20 for example, we train U2MOT for 80 epochs with an extra CrowdHuman [36] dataset. For VisDrone-MOT, U2MOT is trained for 40 epochs without extra datasets. A pre-trained UniTrack [44] ReID model is added for ByteTrack to handle the multi-class multi-object tracking [54]. Specifically, the input image size is 1440×800 for MOT-challenges and 1600×896 for VisDrone-MOT.

Identity labels are unused in ALL training datasets.

4.3. Main Results

MOT-Challanges. Evaluated by the official server, the results on MOT17 and MOT20 benchmarks are illustrated in Tab. 1, which shows U2MOT beats all of the SOTA supervised and unsupervised methods on HOTA and IDF1 metrics. Specifically, it outperforms the SOTA unsupervised UEANet [22] by a large margin (*e.g.*, 1.2% HOTA on MOT17). With the assistance of the ReID head, U2MOT consistently performs better in terms of HOTA and IDF1 against ByteTrack [54]. However, the IDS increases on MOT20, which is mainly because the extracted feature embedding is naturally biased in such severe scenarios. Embedding-based unsupervised methods (including our U2MOT) are inferior to occluded similarities, leading to the IDS increase. Some occlusion-aware optimizations [16, 49] might alleviate this problem. In addition, the MOTA of U2MOT is slightly decreased, which implies that the competition between detection and re-identification tasks should be further explored. Detailed discussions and experiments are provided in Appendix A2.

Dataset	Tracker	*Sup.*	HOTA↑	MOTA↑	IDF1↑	IDS↓
MOT17	TrkFormer [30]	✔	57.3	74.1	68.0	2829
	MOTR [51]	✔	57.8	73.4	68.6	2439
	TraDeS [47]	✔	52.7	69.1	63.9	3555
	CorrTrack [41]	✔	60.7	76.5	73.6	3369
	MTrack [50]	✔	60.5	72.1	73.5	2028
	OUTrack [24]	✗	58.7	73.5	70.2	4122
	PointID [37]	✗	–	74.2	72.4	2748
	UEANet [22]	✗	62.7	77.2	77.0	1533
	ByteTrack [54]	✗	63.1	**80.3**	77.3	2196
	U2MOT (Ours)	✗	**64.2**	79.7	**78.2**	**1506**
MOT20	CorrTrack [41]	✔	–	65.2	69.1	5183
	MTrack [50]	✔	55.3	63.5	69.2	6031
	OUTrack [24]	✗	56.2	68.6	69.4	2223
	UEANet [22]	✗	58.6	73.0	75.6	1423
	ByteTrack [54]	✗	61.3	**77.8**	75.2	**1223**
	U2MOT (Ours)	✗	**62.7**	77.1	**76.2**	1379

Table 1: **Performance comparison against SOTA trackers on MOT-Challenge test sets.** '↑'/'↓' indicates higher/lower values are better, respectively. **Bold** numbers are superior results.

Method	*Sup.*	MOTA↑	IDF1↑	IDS↓	FPS↑
MOTR [51]	✔	22.8	41.4	959	7.5
TrkFormer [30]	✔	24.0	30.5	4840	7.4
UavMOT [25]	✔	36.1	51.0	2775	12.0
ByteTrack [54]	✗	52.3	68.3	1232	11.4
U2MOT (Ours)	✗	**52.3**	**69.0**	1052	**19.4**

Table 2: **Performance on VisDrone-MOT test-dev set.**

In addition, U2MOT does not involve network structure evolution, so the performance gains brought by U2MOT is uncorrelated with those enhancement modules proposed by advanced trackers in Tab. 1. Combining U2MOT with these methods would lead to even better tracking performance.

VisDrone-MOT. For the videos captured in UAV views, the IoU information (or motion model) is unreliable due to the irregular camera motion. To deal with this issue, camera motion compensation [54] and objects' positional relation [25] are mainly adopted, which are effective but computationally expensive. This work provides another solu-

Method	HOTA↑	MOTA↑	IDF1↑	IDS↓
baseline	63.40	73.73	74.51	207
+LTD	63.43	73.74	74.64	202
+UTL	63.84	73.78	75.19	203
+TGA	**64.08**	73.79	**75.42**	**197**
supervised	*63.96*	*73.79*	*75.32*	*196*

Table 3: **Evaluation of the proposed modules.**

TGA-src	TGA-tgt	HOTA↑	MOTA↑	IDF1↑	IDS↓
–	–	63.84	73.78	75.19	203
random	random	63.93	73.79	75.31	204
uncertain	random	63.92	73.77	75.36	194
random	uncertain	64.01	73.79	75.35	201
uncertain	uncertain	**64.08**	73.79	**75.42**	197

Table 4: **Ablation on the anchor-selection mechanism in TGA (Sec. 3.3).** The "–" indicates TGA is not applied.

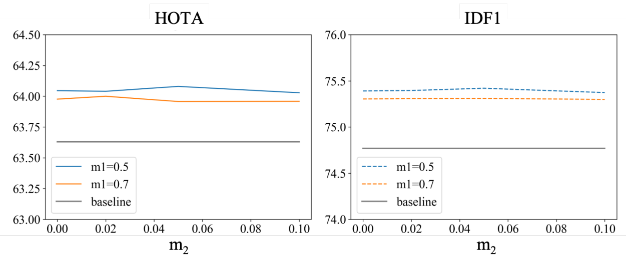


Figure 6: **Ablation on the proposed uncertainty metric.**

tion, *i.e.*, using an uncertainty metric and relaxed motion constraints for refined association results, which is robust to large camera motion as well as low frame rate. As shown in Tab. 2, U2MOT substantially outperforms all the comparison methods, including ByteTrack, who shares the same detector with our U2MOT while utilizes the pre-trained self-supervised UniTrack model [44] for ReID. Also without identity annotation, the IDF1 gains demonstrate our learned task-specific appearance embedding head beats the pre-trained model. Besides, the improved FPS mainly comes from U2MOT's jointly-trained ReID head, rather than an extra ReID model that requires another thorough inference on the raw images. The results demonstrate the effectiveness and efficiency of our proposed U2MOT tracker. Typical tracking visualizations are provided in Appendix J to demonstrate our superiority.

4.4. Ablation Studies

In this section, we conduct extensive ablation studies to elaborate on the effectiveness of the proposed approach. Following the previous methods [55, 22, 47], the first half of each video of the MOT17 training set is used for training, and the second half is for validation. All the models are trained for 30 epochs. Beside the results below, we also conduct ablation studies by training and testing on separate videos with cross-validation. The conclusion is unchanged. For more details please refer to Appendix G.

Effectiveness of the modules proposed in U2MOT. Our unsupervised framework proposes two major components: uncertainty-aware tracklet-labeling (UTL) and tracklet-guided augmentation (TGA). To evaluate each component, we conduct an ablation study on the tracking performance. The results are shown in Tab. 3. We first construct a baseline model by training on adjacent frames. To introduce long-term dependency (LTD), the vanilla similarity-based association on historical frames is conducted for pseudo identities. However, it results in negligible gains in terms of HOTA and MOTA due to the noisy pseudo-labels, meanwhile the IDF1 and IDS obtain slight increases. Instead, the proposed UTL strategy improves the tracking performance in most of the metrics (*e.g.*, 0.4% HOTA), which evidences the fact that long-term temporal consistency is well preserved. Finally, the TGA strategy results in increases of 0.2% HOTA and 0.2% IDF1, as well as decreased IDS, demonstrating that our task-specific augmentation assists in learning the inter-frame consistency. Equipped with the proposed components, unsupervised U2MOT even achieves better HOTA and IDF1 against the identity-supervised model (without UTL and TGA), which validates the effectiveness of our method and indicates the potential to leverage large-scale unlabeled videos.

Uncertain margins. Since the uncertainty metric is vital, we investigate the performance variance caused by different uncertainty margins when verifying the associations. As shown in Fig. 6, different combinations of m_1, m_2 consistently improve the tracking performance. And the improvement is relatively not sensitive to the exact value of these two hyper-parameters. It indicates that wrong associations usually occur in candidates with comparable similarities and relatively lower confidence, which are able to be filtered out and rectified. We choose $m_1 = 0.5, m_2 = 0.05$ for slightly better performance. Moreover, we have provided further experiments on parameter stability with different models in Appendix D, as well as a comparison with other uncertainty metrics in Appendix C and Appendix F.

Augmentation strategies. The customized tracklet-guided augmentation is mainly explored in Tab. 4, where the hierarchical uncertainty-based anchor-sampling mechanism is further evaluated. First, TGA benefits the tracking performance even with totally random anchor-selections. Meanwhile, the selected source anchor tracklet with low-uncertainty avoids dramatic transformation, which brings a slight decrease in IDS. Since most of the pseudo-

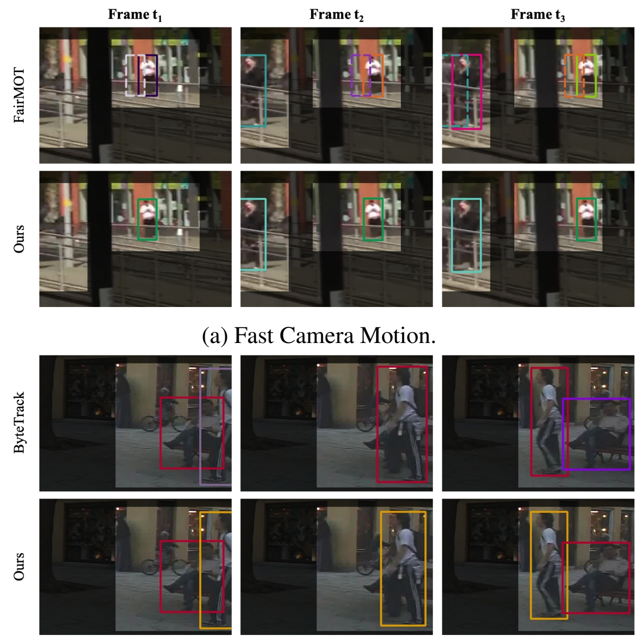
Tracker	HOTA↑	MOTA↑	IDF1↑	IDS↓
U2MOT	64.08	73.79	75.42	**197**
+UTL	**64.90**	**74.09**	**76.66**	212
ByteTrack	63.32	73.72	74.32	**207**
+UTL	**64.90**	**74.09**	**76.66**	212
FairMOT	62.03	72.65	72.83	618
+UTL	**64.01**	**73.35**	**75.05**	**427**
DeepSORT	58.48	70.81	66.20	526
+UTL	**59.75**	**70.97**	**67.60**	**512**
MOTDT	60.49	71.95	69.87	622
+UTL	**61.46**	**72.55**	**71.40**	**353**

Table 5: **Inference boosting.** Results are obtained by different association strategies with the SAME model.

tracklets are accurate enough after the training, this mechanism mostly serves as stabilizing the training in the early stages. Moreover, the selected target anchor object with high-uncertainty along the tracklet brings qualified hard negative examples, leading to a 0.1% HOTA increase. Ultimately, the combined hierarchical uncertainty-based anchor-sampling mechanism results in better performance, demonstrating the effectiveness of TGA. Furthermore, we have quantitatively evaluated the superiority of the TGA strategy over other approaches in Appendix F.

Inference boosting. The proposed uncertainty-aware tracklet labeling (UTL) strategy does not conflict with existing matching strategies. On the contrary, combined with our method, existing methods achieve better tracking performance. As shown in Tab. 5, we first set our method as the comparison baseline, which simply adds ReID embeddings to the ByteTrack [54], and our UTL can thus be equipped. Besides, we integrate UTL to other three popular MOT trackers, including FairMOT [55], DeepSORT [46], and MOTDT [5]. It shows that the UTL consistently boosts all of these trackers by a large margin on most of the metrics, especially in HOTA and IDF1. The IDS of FairMOT and MOTDT is significantly decreased. The training-free UTL shows its effective and generalized application prospects.

Some typical visualization results are shown in Fig. 7, which is consistent with Tab. 5. First, when IoU information is unreliable in irregular camera motions, our method is robust to spatial prediction noise with the uncertainty-based verification stage. Second, in the rectification stage, the tracklet appearance embedding provides important supplementary information to confront the transient occlusions.



(a) Fast Camera Motion.

(b) Object Occlusion.

Figure 7: **Typical visualizations.** (a) Under a moving camera, our method gets rid of the wrong motion prediction from Kalman Filter (dashed boxes). (b) In an occlusion case, tracklet features assist in avoiding ID switches.

5. Methodology Limitation

While U2MOT can enhance the ability of unsupervised trackers by leveraging uncertainty during training, the current implementation has some limitations. One of these limitations is that the uncertainty assessment is conducted offline, which is isolated from the network training process. This means that the model cannot adjust and improve in real-time during training based on the uncertainty analysis, potentially limiting its ability to optimize its performance. Moreover, this offline uncertainty assessment has led to an increase in train time, with the current implementation taking twice as long to train the network. This could be problematic in scenarios where time is a critical factor or when there are large amounts of data to process.

6. Conclusion

This paper proposes a novel unsupervised MOT framework, U2MOT, to address the challenging and underexplored issue of uncertainty in visual tracking. The proposed method improves the quality of pseudo-tracklets through an uncertain-aware tracklet-labeling strategy and enhances tracklet consistency through a tracklets-guide augmentation method that employs a hierarchical uncertainty-based sampling approach for generating hard samples. Experimental results demonstrate the effectiveness of U2MOT , showing the potential of uncertainty. Moving forwards, we will continue to investigate a general video-related uncertainty metric and its applications in various downstream tasks.

References

[1] Yalong Bai, Yifan Yang, Wei Zhang, and Tao Mei. Directional self-supervised learning for heavy image augmentations. In *Proceedings of the IEEE Conference on CVPR*, pages 16692–16701, 2022. 2

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 6

[3] Berta Bescos, Carlos Campos, Juan D Tardós, and José Neira. Dynaslam ii: Tightly-coupled multi-object tracking and slam. *IEEE robotics and automation letters*, 6(3):5191–5198, 2021. 1

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE ICIP*, pages 3464–3468. IEEE, 2016. 1, 2

[5] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *2018 IEEE ICME*, pages 1–6. IEEE, 2018. 8

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 3, 5

[7] 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. 2, 6

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. *arXiv preprint arXiv:1606.03798*, 2016. 5

[9] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 2

[10] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(4):1–18, 2018. 2

[11] Alhussein Fawzi, Horst Samulowitz, Deepak Turaga, and Pascal Frossard. Adaptive data augmentation for image classification. In *2016 IEEE international conference on image processing (ICIP)*, pages 3688–3692. Ieee, 2016. 2

[12] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 6

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on CVPR*, pages 3354–3361. IEEE, 2012. 1

[14] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF CVPR*, pages 2918–2928, 2021. 2

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on CVPR*, pages 9729–9738, 2020. 3

[16] Lingxiao He and Wu Liu. Guided saliency feature learning for person re-identification in crowded scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 357–373. Springer, 2020. 6

[17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*, 2017. 2, 4

[18] Yiqi Jiang, Weihua Chen, Xiuyu Sun, Xiaoyu Shi, Fan Wang, and Hao Li. Exploring the quality of gan generated images for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4146–4155, 2021. 2

[19] Shyamgopal Karthik, Ameya Prabhu, and Vineet Gandhi. Simple unsupervised multi-object tracking. *arXiv preprint arXiv:2006.02609*, 2020. 1, 2

[20] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[21] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[22] 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. 1, 2, 6, 7

[23] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8738–8745, 2019. 2

[24] Qiankun Liu, Dongdong Chen, Qi Chu, Lu Yuan, Lei Zhang, and Nenghai Yu. Online multi-object tracking with unsupervised re-identification learning and occlusion estimation. *Neurocomputing*, 483:333–347, 2022. 1, 2, 3, 6

[25] Shuai Liu, Xin Li, Huchuan Lu, and You He. Multi-object tracking meets moving uav. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8876–8885, 2022. 6

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

[27] Zirui Liu, Haifeng Jin, Ting-Hsiang Wang, Kaixiong Zhou, and Xia Hu. Divaug: Plug-in automated data augmentation with explicit diversity maximization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4762–4770, 2021. 2

[28] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021. 6

[29] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[30] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8844–8854, 2022. 6

[31] 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. 1, 2, 6

[32] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160. IEEE, 2011. 1

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[34] Tarek Said, Samy Ghoniemy, and Omar Karam. Real-time multi-object detection and tracking for autonomous robots in uncontrolled environments. In *2012 Seventh ICCES*, pages 67–72. IEEE, 2012. 1

[35] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *NeurIPS*, 31, 2018. 2

[36] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 6

[37] Bing Shuai, Xinyu Li, Kaustav Kundu, and Joseph Tighe. Id-free person similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14689–14699, 2022. 1, 2, 3, 6

[38] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on CVPR*, pages 2446–2454, 2020. 1

[39] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? *NeurIPS*, 34:29074–29087, 2021. 2

[40] Hao Wang, Qilong Wang, Fan Yang, Weiqi Zhang, and Wangmeng Zuo. Data augmentation for object detection via progressive and selective instance-switching. *arXiv preprint arXiv:1906.00358*, 2019. 2

[41] Qiang Wang, Yun Zheng, Pan Pan, and Yinghui Xu. Multiple object tracking with correlation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3876–3886, 2021. 6

[42] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019. 2

[43] Zhongdao Wang, Jingwei Zhang, Liang Zheng, Yixuan Liu, Yifan Sun, Yali Li, and Shengjin Wang. Cycas: Self-supervised cycle association for learning re-identifiable descriptions. In *European Conference on Computer Vision*, pages 72–88. Springer, 2020. 1, 2

[44] 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. 6, 7

[45] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2019. 2

[46] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE ICIP*, pages 3645–3649. IEEE, 2017. 1, 8

[47] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 12352–12361, 2021. 6, 7

[48] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018. 3

[49] Cheng Yan, Guansong Pang, Jile Jiao, Xiao Bai, Xuetao Feng, and Chunhua Shen. Occluded person re-identification with single-scale global representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11875–11884, 2021. 6

[50] En Yu, Zhuoling Li, and Shoudong Han. Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 8834–8843, 2022. 6

[51] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 6

[52] Fangneng Zhan and Changgong Zhang. Spatial-aware gan for unsupervised person re-identification. In *25th ICPR*, pages 6889–6896. IEEE, 2021. 1, 2

[53] Junbo Zhang and Kaisheng Ma. Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 16650–16659, 2022. 2

[54] 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. 2, 6, 8

[55] 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. 1, 2, 5, 6, 7, 8

[56] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018. 1

[57] 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. 2, 6