Heterogeneous Diversity Driven Active Learning for Multi-Object Tracking

Rui Li[1,2,*], Baopeng Zhang[1], Jun Liu[2], Wei Liu[1], Jian Zhao[3,4,5], Zhu Teng[1,†]

[1]Beijing Jiaotong University [2]Singapore University of Technology and Design
[3]Institute of North Electronic Equipment [4]Peng Cheng Laboratory [5]Intelligent Game and Decision Laboratory

Abstract

The existing one-stage multi-object tracking (MOT) algorithms have achieved satisfactory performance benefiting from a large amount of labeled data. However, acquiring plenty of laborious annotated frames is not practical in real applications. To reduce the cost of human annotations, we propose Heterogeneous Diversity driven Active Multi-Object Tracking (HD-AMOT), to infer the most informative frames for any MOT tracker by observing the heterogeneous cues of samples. HD-AMOT defines the diversified informative representation by encoding the geometric and semantic information, and formulates the frame inference strategy as a Markov decision process to learn an optimal sampling policy based on the designed informative representation. Specifically, HD-AMOT consists of a diversified informative representation module as well as an informative frame selection network. The former produces the signal characterizing the diversity and distribution of frames, and the latter receives the signal and conducts multi-frame cooperation to enable batch frame sampling. Extensive experiments conducted on the MOT15, MOT17, MOT20, and Dancetrack datasets demonstrate the efficacy and effectiveness of HD-AMOT. Experiments show that under 50% budget our HD-AMOT can achieve similar or even higher performance as fully-supervised learning.

1. Introduction

As a sophisticated research topic, one-stage multi-object tracking (MOT) has attracted more research attention in recent years, which simultaneously focuses on pinpointing multiple objects in each frame and recording their trajectories in the entire sequence. Despite the great success of deep learning-based one-stage MOT methods [25, 48, 27, 26, 52, 6, 22], they need a large amount of annotated samples for training, which requires costly human labor. For example, in the MOT17 dataset [23], a sequence containing only 600 frames requires annotating 18,581 bounding boxes (including coordinates and identity labels) for 62 pedestrians. In addition, there are many similar continuous frames in tracking sequences, such as frames 365-368 of the MOT17-02 sequence in Fig. 1, which may have little effect on boosting the performance of the tracker. One way to deal with the above issue is to randomly or uniformly sample tracking sequences for annotation. However, as shown in the red dotted box of Fig. 1, objects in tracking sequences may have different motion states, and the moving state contains more knowledge than the static state, in which case the uniform sampling may fail to capture diverse visual patterns. Moreover, random sampling is easy to yield unstable results. Therefore, it is essential to develop algorithms to infer training data more efficiently in one-stage MOT.

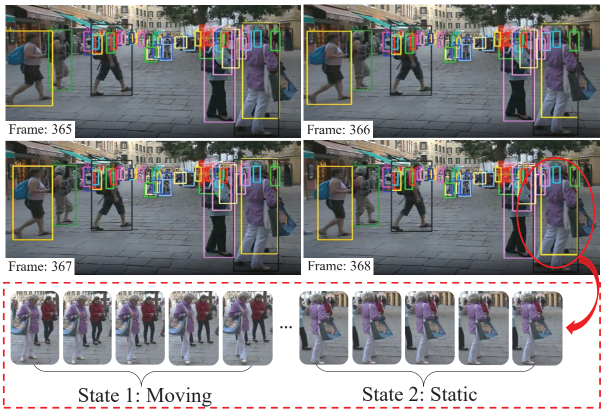

Figure 1. A stimulus example of active learning for one-stage MOT. There are many similar continuous frames in tracking sequences, such as frames 365-368 of MOT17-02 [23], and fully annotated these frames may have little effect on improving the performance of the tracker. Moreover, objects usually have different states, such as moving and static states, as shown in the red dotted box. The moving states usually contain more knowledge than the static ones, in which case the uniform and random sampling may fail to capture the diverse visual patterns of tracking sequences.

Instead of selecting the subset randomly or uniformly,

[*] Work done when Rui Li was an visiting student of SUTD.

[†] Zhu Teng is the corresponding author (zteng@bjtu.edu.cn).

active learning (AL) aims to infer the most informative samples as training data to improve the knowledge of a model, which is a promising way to select training data more efficiently for one-stage MOT. In the computer vision area, active learning has been widely explored for image classification [3, 21, 1, 8], object detection [45, 44, 10], semantic segmentation [1, 7, 28, 9], and so on. In these research topics, the image selection strategies of uncertainty-based [32, 42] and distribution-based methods [31, 1] become the mainstream of active learning. However, these AL algorithms do not directly relate to the performance improvements of the specific task and usually lead to suboptimal performance. More notably, they are usually utilized to select independent image samples, which are prone to achieve poor coverage of the tracking sequence and produce redundant frames if they are employed to the one-stage MOT task stiffly and directly. In particular, since new individuals appear constantly in the tracking sequence, the distribution bias caused by the poor coverage and redundant frames may make the tracker unable to generalize well for unseen cases.

Furthermore, different from other tasks that distinguish inter-class objects in each image (such as image classification, object detection, *etc*.), one-stage MOT needs to detect and associate intra-class objects, *e.g.*, pedestrians belonging to different individuals. This implies that the active learning for one-stage MOT is more challenging, because it should consider not only the discrepancy between frame-level images, but also the influence of more fine-grained intra-class diversity in sample selection, which plays a vital role in object association in one-stage MOT.

To address the above issues, we formulate the active learning of one-stage MOT as a Markov Decision Process (MDP) [36] and design a novel Heterogeneous Diversity driven Active Multi-Object Tracking (HD-AMOT) framework, which leverages a Diversified Informative Representation Module (DIRM) and an Informative Frame Selection Network (IFSN) to learn a frame sampling policy. Specifically, IFSN is designed to infer a batch of informative and diverse frames for annotation by receiving the signal characterizing the diversity and distribution of frames and conducting multi-frame cooperation, which can effectively inhibit poor coverage and reduce redundant frames. In addition, to facilitate the inference of benignant frames, we develop a novel DIRM to characterize the frame-level diversified representation of samples by encoding the geometric and semantic cues of frames and objects. Besides the frame-level diversity representation, we also evaluate the set-level (labeled set and unlabeled set) discrepancy in DIRM, which dedicates to the acquirement of the training set with the unbiased distribution.

Our main contributions are summarized as follows: (1) We propose a novel HD-AMOT for one-stage MOT, which is formulated as a Markov Decision Process (MDP) and learns a sampling policy driven by the reward that directly relates to the tracker performance. To the best of our knowledge, this is the first work that investigates the annotation budget for one-stage MOT based on active learning. (2) To infer the benignant frames, we design a diversified informative representation to characterize the diversity and distribution of frames, which learns both the set-level discrepancy and frame-level diversity by encoding the heterogeneous cues of geometry and semantics. (3) Extensive experiments are conducted on four datasets to demonstrate the efficacy and effectiveness of HD-AMOT. Experiments show that our HD-AMOT can achieve similar or even higher performance as fully-supervised learning under 50% annotation budget.

2. Related Work

Multi-Object Tracking. MOT is a challenging task due to the complexity of tracking scenes. The existing works mainly follow the tracking-by-detection paradigm and two-stage methods are the mainstream, which regards the detection and tracking as two independent models [4, 14, 46, 40, 11, 17]. Despite noticeable progress, the two-stage methods cause a waste of resources and maybe lead to an efficiency issue because the independent detection and tracking models are executed separately in these methods. Recently, with the quick maturity of deep learning[50, 49, 51], one-stage frameworks for MOT [25, 48, 27, 26, 53, 52, 6, 22, 47] have begun to attract more research attention, which greatly improve the accuracy and efficiency of tracking. The fundamental concept revolves around the simultaneous integration of object detection and association within a singular network architecture. For example, CTracker [27] employs a chaining mechanism that connects paired bounding-box regression outcomes derived from overlapping nodes. Each of these nodes spans two consecutive frames. FairMOT [48] presents a simple yet effective approach based on the detection architecture CenterNet [13]. Despite the notable advancements introduced by these methods, their practical applicability remains constrained by the demanding nature of high-cost annotations.

To reduce the demand for labeled data, several methods with less supervision signal such as weakly-supervised and self-supervised learning [30, 2] can effectively utilize the unlabeled data to improve the MOT performance. However, these methods pay more attention to the tracker learning, and ignore the importance of labeled data selection to improve the tracker performance. In this paper, we design an active learning approach for one-stage MOT, which could significantly reduce the annotation cost. To the best of our knowledge, this is the first work that investigates the annotation budget for one-stage MOT based on active learning.

Active Learning. Active learning is an important machine learning problem, which has received lots of attention in computer vision, such as image classification

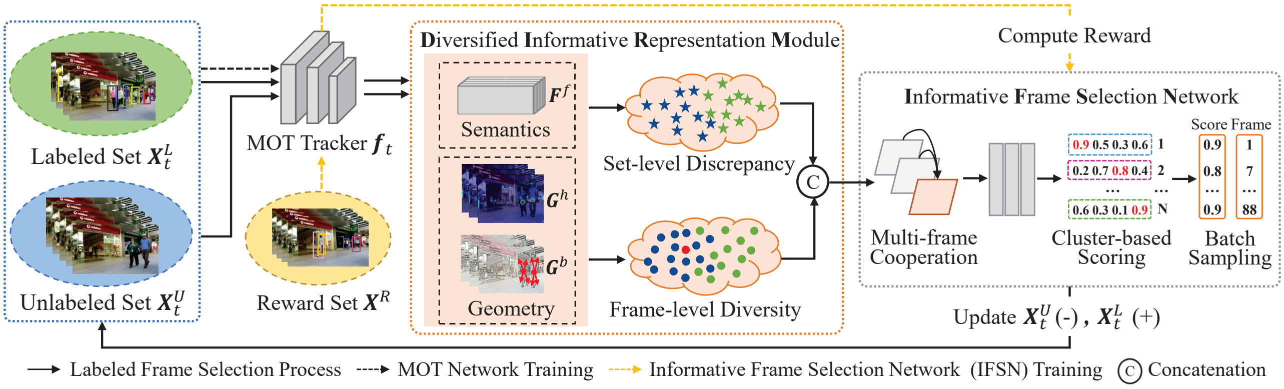

Figure 2. The pipeline of our proposed HD-AMOT framework. In our method, we perform several iterations until a budget $\mathcal{B}$ of labeled frames is achieved. In the t^{th} iteration, the steps are as follows : (I) A one-stage MOT tracker $\boldsymbol{f}_t$ is used to obtain the heterogeneous cues including semantic feature $\boldsymbol{F}^f$ and geometric information $\boldsymbol{G}^h$, $\boldsymbol{G}^b$ on each frame; (II) The set-level discrepancy and frame-level diversity are estimated and construct the MDP state-action pair (s_t, a_t); (III) The multi-frame cooperation based on (s_t, a_t) is performed; (IV) With the assistance of multi-frame cooperation, IFSN infers N unlabeled frames with high scores to be annotated. Meanwhile, $\boldsymbol{X}_t^U$, $\boldsymbol{X}_t^L$ are updated to $\boldsymbol{X}_{t+1}^U$, $\boldsymbol{X}_{t+1}^L$ by displacing newly annotated frames from $\boldsymbol{X}_t^U$ to $\boldsymbol{X}_t^L$; (V) The tracker $\boldsymbol{f}_t$ is retrained on $\boldsymbol{X}_{t+1}^L$ to obtain the updated $\boldsymbol{f}_{t+1}$. (VI) The reward r_{t+1} based on $\boldsymbol{f}_t$ and $\boldsymbol{f}_{t+1}$ is computed on $\boldsymbol{X}^R$ to update parameters of IFSN.

[1, 8, 39, 20], object detection [45, 44, 10], semantic segmentation [7, 28, 33], and so on. These methods can be divided into uncertainty-based [32, 42] and distribution-based methods [31, 1]. For example, [15] exploits the uncertainty in both the input and output spaces to select the most valuable information. [31] defines the problem of active learning as core-set selection. [32] develops a foundation LearningLoss++ for the active learning of pose estimation to establish equivalency between Learning Loss [42] empirically driven objective and the KL divergence objective. These AL algorithms are used to select independent image samples for a specific task yet are not suitable for the video-wise one-stage MOT task. It is prone to cause the problems such as poor coverage and redundant frames if we directly apply the existing AL approaches to one-stage MOT stiffly. Accordingly, we design an active learning framework HD-AMOT for one-stage MOT, which is formulated as a Markov Decision Process and learns a batch frame sampling policy driven by the reward that directly relates to the performance of the MOT tracker, which gives consideration to both sampling coverage and efficiency.

MDP of Multi-Object Tracking. As a sequential decision model, Markov decision process is usually applied in dynamic environments where an agent needs to perform certain tasks by making decisions and executing actions sequentially. In multi-object tracking, several MDP-based works have been explored [41, 29, 16]. [41] formulates the online two-stage MOT problem as decision making in MDPs, where the lifetime of an object is represented using an MDP model. [29] treats individual objects as agents, utilizing a prediction network for tracking, while optimizing tracking outcomes through collaborative interactions among various agents and their environments, facilitated by the decision network. In this paper, different from these methods that employ the MDP to directly solve the detection and tracking subtask of MOT, we address the annotation budget for one-stage MOT based on active learning by the designed Markov decision process, and design a diversified informative representation with an informative frame selection network to enable effective batch sampling under a specific annotation budget.

3. Our Proposed HD-AMOT

3.1. Overview

In this paper, given an unlabeled video sequence $\boldsymbol{X}$ with a limited annotation budget $\mathcal{B}$, our HD-AMOT model aims to learn an optimal sampling strategy that infers and annotates the most informative frames iteratively to maximize the performance of the tracker $\boldsymbol{f}$. Specifically, we cast this AL problem of one-stage MOT by an MDP and adopt the Q-learning algorithm [24] to solve this problem, where we introduce an IFSN to infer frames according to the designed state-action representation (s_t, a_t). As shown in Fig. 2, in our method, the video sequence $\boldsymbol{X}$ is divided into three different subsets $\boldsymbol{X}^U$, $\boldsymbol{X}^L$, and $\boldsymbol{X}^R$. And at each iteration t, the following steps are executed:

(I) The heterogeneous cues including semantic features $\boldsymbol{F}^f$ and geometric information $\boldsymbol{G}^h$, $\boldsymbol{G}^b$ of each frame in the labeled set $\boldsymbol{X}_t^L$ and unlabeled set $\boldsymbol{X}_t^U$ are obtained by $\boldsymbol{f}_t$; (II) We estimate the set-level discrepancy and frame-level diversity in DIRM based on the obtained heterogeneous cues, which are regarded as the MDP state-action pair (s_t, a_t); (III) The multi-frame cooperation based on (s_t, a_t) is performed on $\boldsymbol{X}_t^L$ and $\boldsymbol{X}_t^U$; (IV) With the assistance of multi-frame cooperation, IFSN selects N unlabeled frames

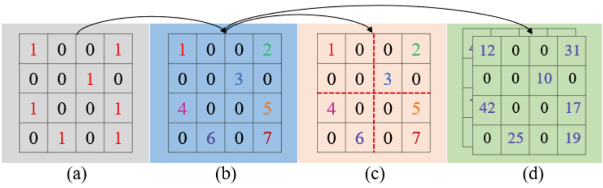

Figure 3. Illustration of geometric matrices construction process. (a): binary heatmap; (b): global topology matrix $\boldsymbol{G}^h$; (c): local topology matrices $\boldsymbol{G}^{h_1} \sim \boldsymbol{G}^{h_K}$; (d): object scale matrix $\boldsymbol{G}^b$.

with high scores to be annotated guided by frame clusters. Meanwhile, the subsets $\boldsymbol{X}_t^U$, $\boldsymbol{X}_t^L$ are updated to $\boldsymbol{X}_{t+1}^U$, $\boldsymbol{X}_{t+1}^L$ by moving newly annotated frames from $\boldsymbol{X}_t^U$ to $\boldsymbol{X}_t^L$; (V) The tracker $\boldsymbol{f}_t$ is retrained on $\boldsymbol{X}_{t+1}^L$ to obtain the updated $\boldsymbol{f}_{t+1}$; (VI) The reward r_{t+1} based on $\boldsymbol{f}_t$ and $\boldsymbol{f}_{t+1}$ is computed on the reward subset $\boldsymbol{X}^R$, which is used to update parameters of IFSN.

3.2. Diversified Informative Representation

To infer the most informative and representative frames, we propose a DIRM to compute the state and action of the designed MDP, which characterizes the diversity and distribution of frames. Specifically, to compensate for the distribution drift between the unlabeled set $\boldsymbol{X}_t^U$ and the labeled set $\boldsymbol{X}_t^L$, we use heterogeneous cues to evaluate the set-level discrepancy as state s_t. Moreover, from the perspective of selecting the unlabeled frame with the potential contribution to the MOT tracker training, we propose to capture the frame-level diversity of each unlabeled frame as action a_t.

Set-level Discrepancy. There are a large number of objects belonging to different individuals in the tracking sequence, and new individuals may appear constantly over time. To make the trained tracker well generalized to unseen individuals, it is necessary to generate an unbiased training set that captures the whole sequence distribution. Based on this intuition, we propose to narrow the distribution gap between the unlabeled and labeled sets by evaluating set-level discrepancy, to ensure the diversity and representativeness of the selected frames.

Specifically, we consider two key attributes to characterize the distribution drifts between the labeled and unlabeled sets: semantic variation and spatial topological variation. We collect the global feature from the last feature extraction layer of $\boldsymbol{f}_t$ to obtain the semantic feature $\boldsymbol{F}^f$, which depicts the general semantic information of the frame. For the spatial topology $\boldsymbol{G}^h$, we encode the center of each object obtained by the tracker $\boldsymbol{f}_t$ into a binary heatmap to represent the spatial distribution of objects. Considering that identities of objects in the frames with similar spatial topology may be quite different due to the object motion, we encode the identities into the binary heatmap to obtain the final $\boldsymbol{G}^h$, as shown in Fig.3 (a) and (b). In addition, the performance of the MOT tracker is affected by objects or background, leading to different tracking quality over various local spatial areas. To suppress adverse effects of the low-quality local areas, we decompose the global spatial topology $\boldsymbol{G}^h$ into K local parts, which form our final spatial topology$\{\boldsymbol{G}^h, \boldsymbol{G}^{h_1-h_K}\}$ together with the global topology. Similarly, the global and local semantic features constitute the final semantics$\{\boldsymbol{F}^f, \boldsymbol{F}^{f_1-f_K}\}$. In this way, the feature space $S_{set} = \{\boldsymbol{G}^h, \boldsymbol{G}^{h_1-h_K}, \boldsymbol{F}^f, \boldsymbol{F}^{f_1-f_K}\}$ is formed. To model the set-level discrepancy between $\boldsymbol{X}_t^L$ and $\boldsymbol{X}_t^U$, Maximum Mean Discrepancy [37] is adopted to calculate the set gap for each feature S_{set}^m in the feature space S_{set} as described in Eq. (1). Finally, the state representation s_t is defined as the concatenation of all $\mathcal{D}^m$ to encode the distribution drifts between $\boldsymbol{X}_t^L$ and $\boldsymbol{X}_t^U$.

$$\begin{aligned}\mathcal{D}^m =& \frac{1}{n_l^2}\sum_{i,j}\varphi(p_i^m, p_j^m) + \frac{1}{n_u^2}\sum_{i,j}\varphi(q_i^m, q_j^m) \\ &- \frac{2}{n_l n_u}\sum_{i,j}\varphi(p_i^m, q_j^m),\end{aligned} \tag{1}$$

where $\mathcal{D}_m$ is a scalar representing the distribution discrepancy between $\boldsymbol{X}_t^L$ and $\boldsymbol{X}_t^U$ on S_{set}^m. p^m and q^m denote the corresponding features of frames in $\boldsymbol{X}_t^L$ and $\boldsymbol{X}_t^U$, n_l and n_u are numbers of frames in $\boldsymbol{X}_t^L$ and $\boldsymbol{X}_t^U$. $\varphi(.)$ is the radial kernel [37] to measure the distance between two features.

Frame-level Diversity. In this paper, we aim to select the frames that are beneficial to the tracker training for annotation. For this purpose, besides considering the set-level discrepancy, we also need to explore frame diversity from different perspectives to infer the most representative and informative frames. Different from the semantic feature and spatial topology used to learn set-level discrepancy, finer-grained cues are desiderated to evaluate the potential contribution of each unlabeled frame. Specifically, besides the frame-wise spatial topology $\boldsymbol{G}^h$ and semantic feature $\boldsymbol{F}^f$, we also consider finer-grained attribute characterizing the individuation of objects: the scale of bounding boxes $\boldsymbol{G}^b$. Its construction process is similar to $\boldsymbol{G}^h$, except that the identities representing different individuals are replaced by the width and height of the corresponding object as shown in Fig.3 (d), *i.e.*, $\boldsymbol{G}^h \in \mathbb{R}^{H\times W\times 1}$ and $\boldsymbol{G}^b \in \mathbb{R}^{H\times W\times 2}$.

Intuitively, we argue that any one of the geometric features $\boldsymbol{G}^h$, $\boldsymbol{G}^b$ and semantic features $\boldsymbol{F}^f$ is the key factor to determine whether a frame contributes to the performance improvement of the tracker. Accordingly, we consider the novelty and prominence of each unlabeled frame x^u based on $S_{frame} = \{\boldsymbol{G}^h, \boldsymbol{G}^b, \boldsymbol{F}^f\}$, to form the frame-level diversity as action a_t. The novelty of an unlabeled frame x_i^u in X_t^U is reflected in the similarities between x_i^u and the labeled frames in $\boldsymbol{X}_t^L$, because the unlabeled frame with a novel characteristic (such as new individuals, different object scales, and postures) usually has low similarities to labeled frames of $\boldsymbol{X}_t^L$. And the prominence of x_i^u lies in the difference between x_i^u and other unlabeled frames in $\boldsymbol{X}_t^U$, so we expect to select the most representative frame by recording the similarity distribution between each unlabeled frame and other frames in $\boldsymbol{X}_t^U$. Specifically, we introduce histogram-based representation to learn the novelty and prominence of each unlabeled frame. Taking x_i^u as an example, to represent the novelty of x_i^u, we calculate cosine similarities between x_i^u and $\boldsymbol{X}_t^L$ on each heterogeneous feature S_{frame}^m and then obtain the histogram $\mathcal{H}_{nov}$ based on cosine similarities. Similar to novelty learning, we learn prominence by recoding the histogram $\mathcal{H}_{pro}$ based on cosine similarities between x_i^u and $\boldsymbol{X}_t^U$. Finally, the action representation a_t of x_i^u is defined as $\{\mathcal{H}_{nov}^{\boldsymbol{G}^h}, \mathcal{H}_{nov}^{\boldsymbol{G}^b}, \mathcal{H}_{nov}^{\boldsymbol{F}^f}, \mathcal{H}_{pro}^{\boldsymbol{G}^h}, \mathcal{H}_{pro}^{\boldsymbol{G}^b}, \mathcal{H}_{pro}^{\boldsymbol{F}^f}\}$, which conduces to IFSN to effectively identify and select the informative and representative frames.

3.3. Informative Frame Selection Learning

Inspired by DQN [38], our lightweight IFSN follows an optimal policy that maximizes the potential performance gains. Considering the sampling efficiency, we propose a batch sampling method guided by frame clusters according to temporal clues. Meanwhile, to restrain the inferior performance caused by redundant frames, we further introduce the multi-frame cooperation strategy before batch sampling.

Multi-frame Cooperation. There are many similar adjacent frames in tracking sequences, which easily leads to redundancy in the process of frame selection. To reduce redundant frames, we introduce a simple multi-frame cooperation strategy to model the neighborhood cooperation between the unlabeled frame x_i^u and its nearest labeled frames. Specifically, we seek the nearest neighbors x_v^l and x_w^l of x_i^u in X_t^L and calculate their actions a_t^v, a_t^w. Then we use the fixed-length compact representation $o_t = \frac{1}{2}(a_t^v + a_t^w)$ as an extra action for x_i^u. To this end, the selection made by IFSN becomes:

$$\phi_t = arg \max_{a_t \in A_t} \; g(s_t, a_t, o_t; \theta), \tag{2}$$

where $a_t \in A_t$ denotes the candidate action in X_t^U, and $g(.)$ is the lightweight IFSN network parameterized by θ.

Batch Frame Selection. The structure of our designed lightweight IFSN consists of four linear layers, where a linear layer parameterized by θ_1 is used to model the relation between each unlabeled frame and its nearest neighbors, and three linear layers parameterized by $\theta_{2\sim4}$ are employed to score each unlabeled frame, as described in Eq. (3), where $cat[.]$ denotes the concatenation operation.

$$g(s_t, a_t, o_t; \theta) = \theta_{2\sim4} \left(cat \left[\theta_1(a_t, o_t), a_t, s_t \right] \right), \tag{3}$$

Intuitively, each unlabeled frame in X_t^U can obtain the corresponding score by IFSN. However, sampling a single frame in each iteration by Eq. (2) to query annotation is inefficient. In this paper, we propose a batch sampling method guided by frame clusters to sample N frames in each iteration. Based on the intuition that a larger difference exists between frames with longer time intervals, we divide the unlabeled frames into N clusters in chronological order and select the frame for annotation in each cluster independently, which is equivalent to labeling a frame in parallel with N annotators. In this case, ϕ_t in Eq. (2) is extended to consist of N independent sub-selections $\{\theta_t^n\}_{n=1}^N$, each with a restricted space A_t^n, as described in Eq. (4).

$$\phi_t^n = arg \max_{a_t \in A_t^n} \; g(s_t, a_t, o_t; \theta), \tag{4}$$

where $a_t \in A_t^n$ denotes the candidate action in n^{th} cluster.

IFSN Optimization. In our method, we compute the reward on a subset X^R, which is used to evaluate the benefit of selected frames to the tracker and optimize the frame selection network IFSN. To evaluate the performance of the tracker more comprehensively, we utilize MOTA [5] and IDF1 [19] as performance metrics to calculate the reward, where MOTA is a comprehensive metric to evaluate the detection performance and IDF1 tends to estimate the tracking ability of a tracker. As described in Eq. (5), the reward r_{t+1} is defined as the difference of the performance metrics between f_{t+1} and f_t.

$$r_t = (e_{t+1}^{mota} + e_{t+1}^{idf1}) - (e_t^{mota} + e_t^{idf1}), \tag{5}$$

where e_{t+1}^{mota} and e_{t+1}^{idf1} are MOTA and IDF1 metrics computed by f_{t+1}, e_t^{mota} and e_t^{idf1} are evaluated by f_t.

With the reward r_{t+1}, we can optimize IFSN to infer the most informative frames to maximize the reward, leading to improved tracker performance during each active learning iteration. More specifically, we train IFSN in a double DQN formulation [38] by optimizing the loss based on the temporal difference error [35] as:

$$\mathcal{L}_{\text{IFSN}} = \left(\hat{y}_t - g(s_t, a_t, o_t; \theta) \right)^2, \tag{6}$$

$$\hat{y}_t = r_{t+1} + \gamma \; g(s_{t+1}, a_{t+1}, o_{t+1}; \hat{\theta}), \tag{7}$$

where θ and $\hat{\theta}$ are parameters of the IFSN policy network and off-policy network following the setting of double DQN [38]. γ denotes a discount factor.

With the described diversified informative representation in Section 3.2 and informative frame selection learning in this section, we introduce our entire HD-AMOT as illustrated in Alg. 1.

4. Experiment

4.1. Experimental Settings

Datasets. We evaluate our HD-AMOT on MOT15 [18], MOT17 [23], MOT20 [12], and Dancetrack [34]. MOT15 and MOT17 are widely-used MOT datasets, and they are

Algorithm 1: HD-AMOT Algorithm.

```
Input: The sequence $\boldsymbol{X}$, initial tracker $f_{init}$ and IFSN
       $g_{init}$, AL procedure rounds $Z$.
$\boldsymbol{X}_{init}$, $\boldsymbol{X}^{R}$ ← RandomPartition ($\boldsymbol{X}$);
$\boldsymbol{X}^{U}_{init}$, $\boldsymbol{X}^{L}_{init}$ ← UniformPartition ($\boldsymbol{X}_{init}$);
$\boldsymbol{X}^{L}_{0} \leftarrow \boldsymbol{X}^{L}_{init}$, $f_0 \leftarrow$ update($g_{init}, \boldsymbol{X}^{L}_{init}$);
while not done do                // Episodes
    for $t = 0$ to $t = Z - 1$ do    // AL procedure
        Learn the diversified representation to form the
          state and action space;
        Use IFSN to select frames by Eq. (4);
        Annotate the selected $N$ frames $\boldsymbol{X}^{S}_{t}$;
        Update $\boldsymbol{X}^{U}_{t}$ ,$\boldsymbol{X}^{L}_{t}$:
        $\boldsymbol{X}^{U}_{t+1} \leftarrow \boldsymbol{X}^{U}_{t} \setminus \boldsymbol{X}^{S}_{t}, \boldsymbol{X}^{L}_{t+1} \leftarrow \boldsymbol{X}^{L}_{t} \cup \boldsymbol{X}^{S}_{t}$;
        $f_{t+1} \leftarrow$ update($f_t, \boldsymbol{X}^{L}_{t+1}$);
        Calculate the reward on $\boldsymbol{X}^{R}$ by Eq. (5);
    end
    Update IFSN following Eq. (6);
    Reset: $\boldsymbol{X}^{L}_{0} \leftarrow \boldsymbol{X}^{L}_{init}$, $f_0 \leftarrow f_{t+1}$;
end
```

challenging because of the large variety in object pose, lighting, viewpoint, *etc*. In contrast to MOT15 and MOT17, MOT20 addresses the challenge of highly congested scenarios, where the pedestrian density can surge to as high as 246 individuals per frame. Dancetrack is a large-scale MOT dataset covering scenarios with low distinguishability of object appearance and diverse non-linear motion patterns.

Evaluation Metrics. To evaluate the performance of one-stage MOT, the standard metrics MOTA [5] and IDF1 [19] are adopted for evaluation, which indicate the detection and association performance of one-stage MOT. Specifically, MOTA combines three error sources including false positives, missed targets, and identity switches, and IDF1 is the ratio of correctly identified detections over the average number of ground-truth and computed detections.

Implementation Details. In this paper, we adopt the popular one-stage MOT method FairMOT [48] as the MOT tracker. Note that the propose algorithm is suitable for any one-stage MOT tracker. The input frames are resized to 1088 × 608. We set the learning rate of our IFSN to 1e-4 and the discount factor γ in Eq. (7) to 0.9, and the number of local features K is set to 9 according to experiments. For each dataset, we randomly select a small sequence in each video at a ratio of 0.1 to form a reward subset X^R. In the remaining frames, a small amount of frames is uniformly sampled as the initially labeled subset X^L_{init} to initialize the MOT tracker f, and the size is set to 100. For the semantic information, we extract the feature map from the last convolutional layer of ResNet34 as the global semantic feature $\boldsymbol{F}^f \in \mathbb{R}^{512\times19\times34}$. For the geometric clues, we utilize the center coordinates and identities of bounding boxes obtained by the MOT tracker to construct the global spatial topology $\boldsymbol{G}^h \in \mathbb{R}^{152\times272}$, and generate the object scale matrix $\boldsymbol{G}^b \in \mathbb{R}^{152\times272\times2}$ by the box size. In addition, the number of clusters N in the batch sampling is set to 12 for MOT15, MOT17, and MOT20, and is set to 24 for Dancetrack as it is much larger than other datasets.

Table 1. Ablation study on the design of the diversified informative representation.

Representation	Setting	MOTA (%)	IDF1 (%)
Set-level Discrepancy	State *w/o* $\boldsymbol{G}^{h,h1-h9}$	61.73	64.95
	State *w/o* $\boldsymbol{F}^{f,f1-f9}$	61.49	64.12
	State *w/o* $\boldsymbol{G}^{h}$, $\boldsymbol{F}^{f}$	62.74	65.21
	State *w/o* $\boldsymbol{G}^{h1-h9}$, $\boldsymbol{F}^{f1-f9}$	62.92	64.55
Frame-level Diversity	Action *w/o* $\mathcal{H}^{\boldsymbol{G}^h}$	62.83	**66.61**
	Action *w/o* $\mathcal{H}^{\boldsymbol{G}^b}$	62.49	66.23
	Action *w/o* $\mathcal{H}^{\boldsymbol{F}^f}$	61.78	64.95
	Action *w/o* $\mathcal{H}^{*}_{nov}$	61.25	63.41
	Action *w/o* $\mathcal{H}^{*}_{pro}$	62.28	64.16
All	HD-AMOT	**63.23**	66.47

4.2. Ablation Study

In this section, to analyze the impacts of different components, we conduct ablation studies on the popular dataset MOT17 following the division strategy in [43]. We run our experiments 5 times and report the mean performance for all ablation studies.

Effect of Diversified Informative Representation. We perform an ablation study with a 20% annotation budget to evaluate the contribution of our diversified informative representation. As IFSN relies on the state (set-level discrepancy) to decide the frame sampling policy, we investigate the influence of the set-level discrepancy by removing different components individually. Moreover, we also analyze the effect of the action design (frame-level diversity). As shown in Tab. 1, our complete diversified informative representation gives the best performance in terms of MOTA and is only slightly inferior to *Action w/o* $\mathcal{H}^{\boldsymbol{G}^h}$ by 0.14% on IDF1. We speculate that the scale information in $\mathcal{H}^{\boldsymbol{G}^b}$ implicitly indicates the identity and topology clues of $\mathcal{H}^{\boldsymbol{G}^h}$, so the introduction of $\mathcal{H}^{\boldsymbol{G}^b}$ weakens the role of $\mathcal{H}^{\boldsymbol{G}^h}$, making the MOTA of *Action w/o* $\mathcal{H}^{\boldsymbol{G}^h}$ close to HD-AMOT and the IDF1 slightly higher. Furthermore, removing the components of set-level discrepancy or frame-level diversity in the diversified informative representation degrades the tracking performance. The largest decrease of MOTA and IDF1 occurs when the score of maximum similarity with labeled frames *Action w/o* $\mathcal{H}^{*}_{nov}$ is removed, which further verifies the effectiveness of heterogeneous cues evaluating the novel characteristic of unlabeled frames.

Effect of Informative Frame Selection Network. We further validate the performance of our designed IFSN in Tab. 2. We first consider performing the cluster-based batch

Table 2. Ablation study on the different settings of the Informative Frame Selection Network.

Model	Multi-frame Cooperation	Batch Sampling		MOTA (%)↑	IDF1 (%)↑
		Bat.	Clu.		
A	×	✓	✓	62.76	65.52
B	✓	×	×	63.07	65.62
C	✓	✓	×	62.18	64.55
D	✓	✓	✓	**63.23**	**66.47**

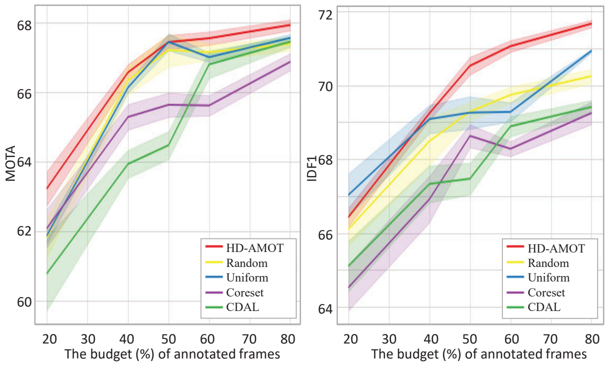

Figure 4. Performance comparison evaluated by MOTA and IDF1 metrics with increasing budgets.

sampling without multi-frame cooperation (*Model A*). Then we analyze the usage of selecting a single frame in each AL iteration (*Model B*). In addition, we also construct the *Model C* as IFSN to select a batch of frames in one shot, where the frames with N highest scores are sampled. Finally, we present the performance of our cluster-based batch sampling strategy (*Model D*). By comparing the results of *Model A*, *Model D* brings performance gains of 0.47% and 0.95% in terms of MOTA and IDF1 metrics, which proves that multi-frame cooperation plays a positive role in the active learning of one-stage MOT. In contrast, *Model C* gives the worst performance, which selects multiple frames according to the N highest scores. We argue that frames with higher scores may be similar samples and close to each other, which easily leads to inefficiency in batch sampling. To this end, cluster-based batch sampling is introduced, as it can avoid selecting adjacent frames with higher scores and learn to sample with better coverage of the underlying distribution. Notably, selecting a single frame at each iteration in *Model B* also provides a competitive performance but still tends to be inferior to our method *Model D*, and the time cost of selecting a single frame per iteration is much higher than our cluster-based batch sampling strategy.

Analysis of Different Budgets. We report the performance of our method, random sampling, uniform sampling, and other existing AL methods such as Coreset [31], CDAL [1] under different budgets, as shown in Fig. 4. For the IDF1 metric, ours is slightly inferior to the uniform sampling when the budget is 20% and 40%. We argue that uniform sampling can select informative frames for annotation if a small number of labeled frames is required. However, when the budget increases, there may be redundancy in the newly acquired frames by uniform sampling, and training with these labels does not provide more additional information, resulting in the stagnation or even decline of the performance of IDF1. Notably, our HD-AMOT with the 50% budget achieves similar performance as other methods with the 80% budget. It reveals that ours can save around 30% labeling efforts on average compared to other sampling methods. We attribute this to our learned diversified informative clues and fine-grained object representation, which can effectively improve the tracking performance.

4.3. Comparison with other AL Methods

To further verify the generalization and effectiveness, we evaluate the proposed framework HD-AMOT on four diversified MOT datasets in this section.

Results on MOT15. Tab. 3 lists the mean values of MOTA and IDF1 on MOT15, and Fig. 5 reveals the corresponding standard deviation of MOTA and IDF1. We can observe that HD-AMOT outperforms all methods by a clear margin, *i.e.*, AL methods used in other tasks (CDAL [1] and Coreset [31]) stiffly applied to one-stage MOT can lead to obvious performance degradation. Notably, although Coreset achieves the second-best performance on MOTA and IDF1, it is clear from the standard deviations in Fig. 5 that Coreset has limited stability. Moreover, the performance of Coreset on other datasets (MOT17, MOT20, and Dancetrack) is not ideal, which indicates that the direct application of Coreset in one-stage MOT is unstable.

Results on MOT17. MOT17 is a challenging dataset with more crowded scenarios, different viewpoints, camera motions, and weather conditions. As shown in Tab. 3, although uniform sampling achieves the best IDF1 on MOT17 and our HD-AMOT performs slightly lower than it by a margin of 0.18%, uniform sampling may fail to capture diverse visual patterns in other datasets due to the lack of flexibility and adaptability. For example, in MOT15, ours outperforms uniform sampling by all performance metrics, which reflects the generalization ability of ours for one-stage MOT. Moreover, besides object association, the object detection ability should be considered in one-stage MOT. Our method, which benefits from learning the sampling policy driven by tracking performance metrics directly from data, significantly outperforms uniform sampling by 0.84% in terms of the detection accuracy metric MOTA.

Results on MOT20. We also use the large-scale dataset MOT20 to demonstrate the scalability of our HD-AMOT as shown in Tab. 3. In contrast to the aforementioned datasets, MOT20 is a recently released dataset that features more densely populated scenes. We can observe that ours still achieves the best performance, which demonstrates that our HD-AMOT can infer informative frames even on a crowded and complicated dataset.

Table 3. Mean values of MOTA and IDF1 metrics for different active learning methods on the test sets of four datasets including MOT15, MOT17, MOT20, and Dancetrack. Best and second-best results are shown in red and **black**, respectively.

Dataset	Method	MOTA(%) (↑)	IDF1(%) (↑)
MOT15	Tracker *w/* Random	53.28	59.61
	Tracker *w/* Uniform	54.09	59.50
	Tracker *w/* CDAL [1]	53.13	58.85
	Tracker *w/* Coreset [31]	**54.70**	**60.29**
	Tracker *w/* **HD-AMOT**	55.29	60.64
MOT17	Tracker *w/* Random	70.11	69.24
	Tracker *w/* Uniform	**70.40**	70.05
	Tracker *w/* CDAL [1]	69.10	68.71
	Tracker *w/* Coreset [31]	69.44	68.61
	Tracker *w/* **HD-AMOT**	71.24	**69.87**
MOT20	Tracker *w/* Random	58.21	65.37
	Tracker *w/* Uniform	**58.65**	**65.87**
	Tracker *w/* CDAL [1]	57.59	65.33
	Tracker *w/* Coreset [31]	56.66	64.86
	Tracker *w/* **HD-AMOT**	59.24	66.41
Dancetrack	Tracker *w/* Random	80.49	40.67
	Tracker *w/* Uniform	81.11	41.03
	Tracker *w/* CDAL [1]	**81.36**	**41.16**
	Tracker *w/* Coreset [31]	81.13	40.47
	Tracker *w/* **HD-AMOT**	81.94	41.93

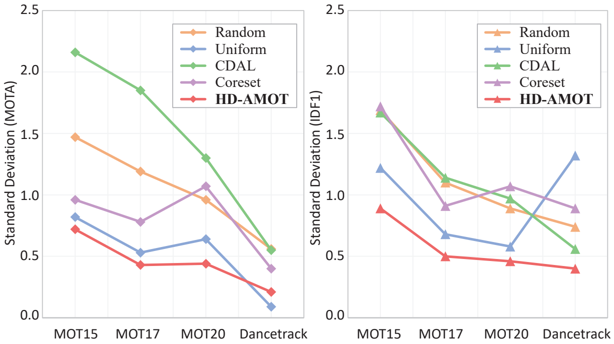

Figure 5. Standard deviations of MOTA and IDF1 metrics for different active learning methods on the test sets of four datasets including MOT15, MOT17, MOT20, and Dancetrack.

Results on Dancetrack. Dancetrack is a larger and more challenging dataset for group dancing scenes, where humans have similar appearance, diverse motion, and extreme articulation. We provide the performance evaluation results on the large-scale Dancetrack in Tab. 3. It can be observed that our HD-AMOT still achieves the best MOTA and IDF1 performance. Notably, from the perspective of sampling stability in Fig. 5, although uniform sampling can obtain a low standard deviation of MOTA, it leads to a very high standard deviation of IDF1 due to failure to capture informative frames that are helpful to object association. By contrast, our proposed HD-AMOT can maintain stable MOTA and IDF1 performance in multiple samplings.

Furthermore, we give the performance comparison between our HD-AMOT with a 50% budget and full supervision in Fig. 6. Our active learning framework achieves similar or even higher performance compared with the tracker with fully-supervised training. For instance, we exceed the full supervision by a gain of 2% on MOTA and 0.8% on IDF1 for MOT15. These results implicitly prove that there are many similar frames in MOT sequences, and these similar frames bring limited progress in tracking performance. Our proposed HD-AMOT learns an optimal sampling policy by the heterogeneous diversity representation and a Markov decision process, which can infer informative frames that can most benefit the tracker, thus effectively improving the performance of the tracker. In summary, the proposed method can provide a consistent gain for the MOT tracker across datasets.

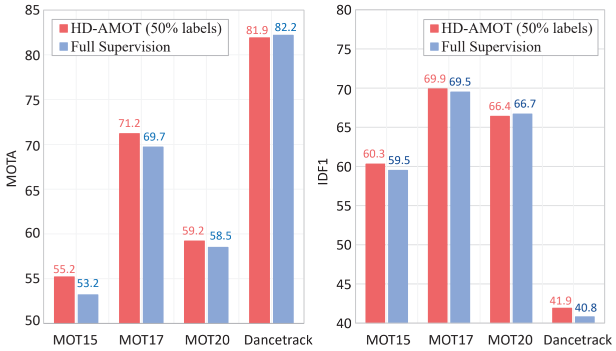

Figure 6. Performance comparison between our HD-AMOT with a 50% budget and full supervision on MOT15, MOT17, MOT20, and Dancetrack datasets.

5. Conclusion

In this paper, we propose a novel framework HD-AMOT to investigate the annotation budget for one-stage MOT. It is formulated as an MDP and learns a frame sampling policy driven by the reward that directly relates to the performance of the tracker. HD-AMOT learns the diversified informative representation of frames and achieves effective cluster-based batch frame selection guided by multi-frame cooperation. We conduct extensive ablation studies to verify the design of our framework and compare our performance with random, uniform sampling, and several existing active learning works on four datasets including MOT15, MOT17, MOT20, and Dancetrack. The experimental results demonstrate the efficacy and effectiveness of our HD-AMOT.

Acknowledgement. This work was supported by the Fundamental Research Funds for the Central Universities of China (2022JBMC009), the Natural Science Foundation of China (61972027, 62006244), and the Beijing Municipal Natural Science Foundation (Grant No. 4212041). This work is also partially supported by Young Elite Scientist Sponsorship Program of China Association for Science and Technology (YESS20200140), and Young Elite Scientist Sponsorship Program of Beijing Association for Science and Technology (BYESS2021178).

References

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *ECCV*, 2020.

[2] Favyen Bastani, Songtao He, and Samuel Madden. Self-supervised multi-object tracking with cross-input consistency. In *NeurIPS*, 2021.

[3] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *CVPR*, 2018.

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

[5] 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] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. In *CVPR*, 2022.

[7] Lile Cai, Xun Xu, Jun Hao Liew, and Chuan Sheng Foo. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *CVPR*, 2021.

[8] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *CVPR*, 2021.

[9] Arantxa Casanova, Pedro O. Pinheiro, Negar Rostamzadeh, and Christopher J. Pal. Reinforced active learning for image segmentation. In *ICLR*, 2020.

[10] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M Alvarez. Active learning for deep object detection via probabilistic modeling. In *CVPR*, 2021.

[11] Peng Dai, Renliang Weng, Wongun Choi, Changshui Zhang, Zhangping He, and Wei Ding. Learning a proposal classifier for multiple object tracking. In *CVPR*, 2021.

[12] 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:2003.09003*, 2020.

[13] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, 2019.

[14] Jiawei He, Zehao Huang, Naiyan Wang, and Zhaoxiang Zhang. Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking. In *CVPR*, 2021.

[15] Sheng-Jun Huang, Nengneng Gao, and Songcan Chen. Multi-instance multi-label active learning. In *IJCAI*, 2017.

[16] Ming-xin Jiang, Chao Deng, Zhi-geng Pan, Lan-fang Wang, and Xing Sun. Multiobject tracking in videos based on lstm and deep reinforcement learning. *Complexity*, 2018, 2018.

[17] Chanho Kim, Li Fuxin, Mazen Alotaibi, and James M Rehg. Discriminative appearance modeling with multi-track pooling for real-time multi-object tracking. In *CVPR*, 2021.

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

[19] Yuan Li, Chang Huang, and Ram Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *CVPR*, 2009.

[20] Zhuoming Liu, Hao Ding, Huaping Zhong, Weijia Li, Jifeng Dai, and Conghui He. Influence selection for active learning. In *ICCV*, 2021.

[21] Zhao-Yang Liu and Sheng-Jun Huang. Active sampling for open-set classification without initial annotation. In *AAAI*, 2019.

[22] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, 2022.

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

[24] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. In *NeurIPS*, 2013.

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

[26] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *CVPR*, 2021.

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

[28] Yu Qiao, Jincheng Zhu, Chengjiang Long, Zeyao Zhang, Yuxin Wang, Zhenjun Du, and Xin Yang. Cpral: Collaborative panoptic-regional active learning for semantic segmentation. In *AAAI*, 2022.

[29] Liangliang Ren, Jiwen Lu, Zifeng Wang, Qi Tian, and Jie Zhou. Collaborative deep reinforcement learning for multi-object tracking. In *ECCV*, 2018.

[30] Idoia Ruiz, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Joan Serrat. Weakly supervised multi-object tracking and segmentation. In *WACV*, 2021.

[31] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018.

[32] Megh Shukla and Shuaib Ahmed. A mathematical analysis of learning loss for active learning in regression. In *CVPR*, 2021.

[33] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *CVPR*, 2020.

[34] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*, 2022.

[35] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[36] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[37] Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. In *NeurIPS*, 2016.

[38] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI*, 2016.

[39] Ran Wang, Sam Kwong, Xu Wang, and Yuheng Jia. Active k-labelsets ensemble for multi-label classification. *PR*, 109:107583, 2021.

[40] Jun Xiang, Guohan Xu, Chao Ma, and Jianhua Hou. End-to-end learning deep crf models for multi-object tracking deep crf models. *IEEE TCSVT*, 31(1):275–288, 2020.

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

[42] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *CVPR*, 2019.

[43] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan. Poi: Multiple object tracking with high performance detection and appearance feature. In *ECCV*, 2016.

[44] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *CVPR*, 2021.

[45] Xiaosong Zhang, Fang Wan, Chang Liu, Xiangyang Ji, and Qixiang Ye. Learning to match anchors for visual object detection. *IEEE TPAMI*, 44(6):3096–3109, 2021.

[46] Yang Zhang, Hao Sheng, Yubin Wu, Shuai Wang, Weifeng Lyu, Wei Ke, and Zhang Xiong. Long-term tracking with deep tracklet association. *IEEE TIP*, 29:6694–6706, 2020.

[47] 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 *ECCV*, 2022.

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

[49] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. Towards pose invariant face recognition in the wild. In *CVPR*, pages 2207–2216, 2018.

[50] Jian Zhao, Lin Xiong, Jianshu Li, Junliang Xing, Shuicheng Yan, and Jiashi Feng. 3d-aided dual-agent gans for unconstrained face recognition. *IEEE TPAMI*, 41(10):2380–2394, 2018.

[51] Jian Zhao, Shuicheng Yan, and Jiashi Feng. Towards age-invariant face recognition. *IEEE TPAMI*, 44(1):474–487, 2020.

[52] Linyu Zheng, Ming Tang, Yingying Chen, Guibo Zhu, Jinqiao Wang, and Hanqing Lu. Improving multiple object tracking with single object tracking. In *CVPR*, pages 2453–2462, 2021.

[53] Xue-Feng Zhu, Tianyang Xu, Jian Zhao, Jia-Wei Liu, Kai Wang, Gang Wang, Jianan Li, Zhihao Zhang, Qiang Wang, Lei Jin, et al. Evidential detection and tracking collaboration: New problem, benchmark and algorithm for robust anti-uav system. *arXiv preprint arXiv:2306.15767*, 2023.