NetTrack: Tracking Highly Dynamic Objects with a Net

Supplementary Material

Algorithm 1: Pseudo code of fine-grained Net.

```
Input: Video sequence $\mathbf{V}$, detector $\mathbb{D}$, physical point tracker
       $\mathtt{Tr}_\mathrm{p}$, tracking stride $\mathcal{S}$
Output: Object tracks $\mathcal{T}_\mathrm{o}$
$\mathcal{T}_\mathrm{o}^\mathrm{coarse}, \mathcal{T}_\mathrm{o}^\mathrm{fine} \leftarrow \emptyset, \emptyset$ # Initialized object tracks
for $\mathbf{I}_k$ in $\mathbf{V}$ do
    $\mathcal{D}_k \leftarrow \mathbb{D}(\mathbf{I}_k)$
    $\mathcal{T}_\mathrm{o}^\mathrm{coarse} \leftarrow \mathtt{Tr}_\mathrm{o}^\mathrm{coarse}(\mathcal{D}_k, \mathcal{T}_\mathrm{o}^\mathrm{coarse})$
    $\mathbf{I}_\mathrm{s}, \mathbf{D}_\mathrm{s} \leftarrow \mathbf{I}_\mathrm{s} \cup \{\mathrm{I}_k\}, \mathbf{D}_\mathrm{s} \cup \{\mathcal{D}_k\}$
    $N \leftarrow$ number of images in $\mathbf{I}_\mathrm{s}$
    if $N == \mathcal{S}$ then
        /* Step 1: Fine-grained sampling */
        $\mathcal{Q} \leftarrow \mathtt{Sampling}(\mathcal{T}_o^\mathrm{coarse})$
        $\mathcal{T}_\mathrm{p} \leftarrow \mathtt{Tr}_\mathrm{p}(\mathbf{I}_\mathrm{s}, \mathcal{Q})$
        /* Step 2: Fine-grained matching */
        for $\mathcal{D}_i$ in $\mathbf{D}_\mathrm{s}$ do
            $\mathcal{T}_\mathrm{o}^\mathrm{fine} \leftarrow \mathtt{matching}(\mathcal{D}_i, \mathcal{T}_\mathrm{o}^\mathrm{fine}, \mathcal{T}_\mathrm{p})$
        end
        $\mathbf{I}_\mathrm{s}, \mathbf{D}_\mathrm{s} = \{\mathrm{I}_k\}, \{\mathrm{D}_k\}$
        $\mathcal{T}_\mathrm{o}^\mathrm{coarse} \leftarrow \mathcal{T}_\mathrm{o}^\mathrm{fine}$
        $\mathcal{T}_\mathrm{o} \leftarrow \mathcal{T}_\mathrm{o}^\mathrm{fine}$
    end
end
Return: $\mathcal{T}_\mathrm{o}$
```

In the supplementary materials, we provide a detailed exposition of the NetTrack pipeline in Sec. 7, statistics and elaborate dynamicity descriptions of the proposed BFT benchmark in Sec. 8, as well as a more comprehensive introduction to experimental details and more enriched experimental results in Sec. 9. Additionally, we also demonstrate three potential application scenarios of NetTrack in Sec. 10.

7. Method Details

In contrast to previous methods [4, 7, 71] that utilize coarse-grained representations, NetTrack introduces a more fine-grained Net by incorporating fine-grained object information for tracking. The brief workflow of the proposed fine-grained Net is illustrated in Algorithm 1. Specifically, after initialization, NetTrack employs a simple and fast coarse tracker to predict the coarse-grained object trajectory $\mathcal{T}_\mathrm{o}^\mathrm{coarse}$. Subsequently, when the frame count reaches a tracking stride $\mathcal{S}$, NetTrack performs fine-grained sampling within this stride, obtaining the points of interest (POIs) $\mathcal{Q}$ and their corresponding trajectories. In fine-grained tracking, the fine-grained object trajectory is obtained by matching with the trajectories of POIs and the coarse-grained object trajectory, which is then used as the output.

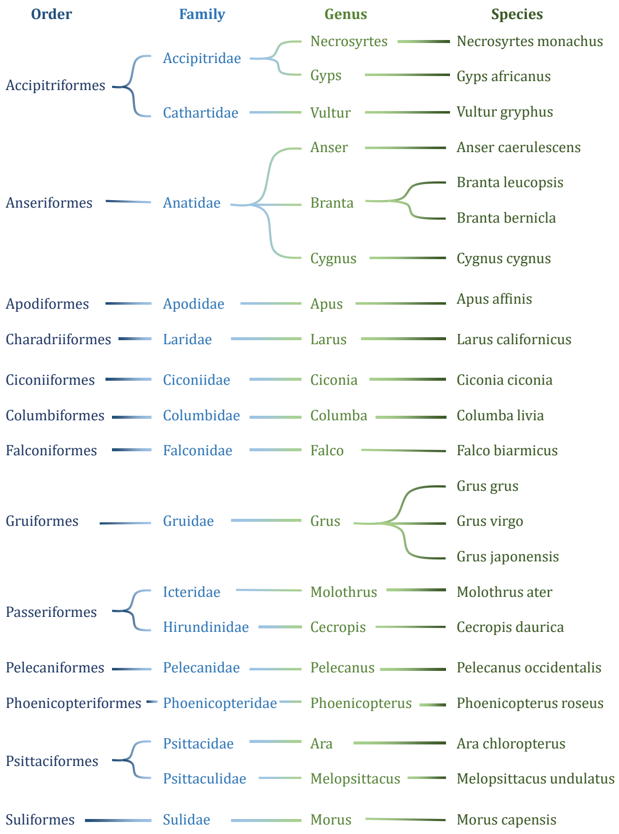

Figure 7. Orders, families, genera, and species of objects in the BFT dataset, showcasing the diversity of the dataset.

8. Benchmark Details

8.1. Statistics

Fig. 1-c and Fig. 4 illustrate the geographical distribution of diverse scenes and environments in the dataset. The varied categories of objects in the BFT dataset also contribute to its diversity, as demonstrated in Fig. 7, which includes 13 orders, 16 families, 19 genera, and 22 species.

8.2. Dynamicity

The dynamicity comparison between BFT and some open-world MOT datasets is shown in Fig. 4-b,c, Fig. 8 further incorporates a comparison with closed-set datasets [9, 49] and involves additional attributes to validate the dynamicity of open-world MOT compared to closed-set datasets, with particular emphasis on the pronounced dynamicity in BFT. All vertical axes in Fig. 8 represent frequency.

IOU The Intersection over Union (IOU) of the same object in adjacent frames can reflect the degree of geometric

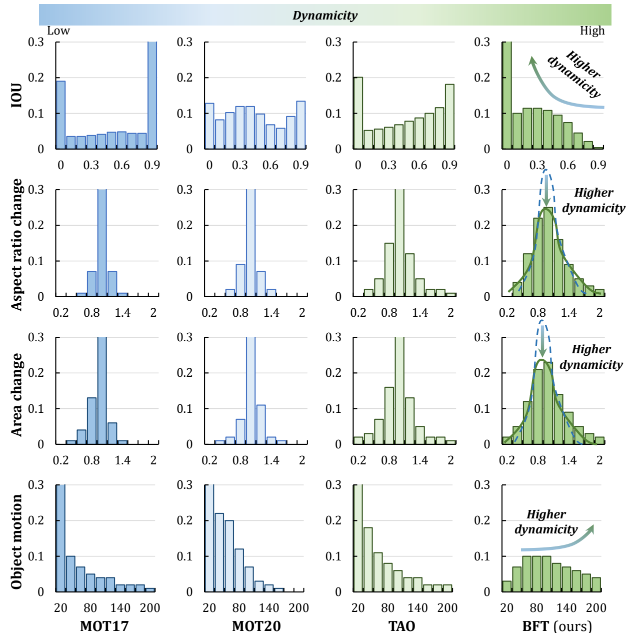

Figure 8. Dynamicity comparison between BFT and other datasets on 4 attributes. BFT exhibits stronger dynamicity.

changes of the object, thus describing its dynamicity. Adjacent IOU is calculated by $\text{IOU}(\mathbf{b}^t, \mathbf{b}^{t-1}) = \frac{|\mathbf{b}^t \cap \mathbf{b}^{t-1}|}{|\mathbf{b}^t \cup \mathbf{b}^{t-1}|}$, where $\mathbf{b}$ is the bounding box of an object, and t denotes frame index. In the first line of Fig. 8, the interval of IOU is 0.1. The lower IOU of BFT indicates its stronger dynamicity in the comparison.

Aspect ratio change Aspect ratio change (ARC) of the same object across adjacent frames reflects the object deformation in the horizontal and vertical directions, which is formulated as $\text{ARC}(\mathbf{b}^t, \mathbf{b}^{t-1}) = \frac{w^{t-1}/h^{t-1}}{w^t/h^t}$. w and h are the width and height of the bounding boxes, respectively. In the second line of Fig. 8, the interval of ARC is 0.2. The further the ARC deviates from 1, the more pronounced the dynamicity of the objects, indicating a greater significance of BFT's dynamicity.

Area change Area change (AC) of the object boxes between adjacent frames reflects the overall size variation of the object. AC is calculated as $\text{AC}(\mathbf{b}^t, \mathbf{b}^{t-1}) = \frac{w^{t-1}h^{t-1}}{w^t h^t}$. In the third line of Fig. 8, the interval of AC is 0.2. BFT's dynamicity is more significant as AC deviates further from 1, resulting in a more pronounced dynamicity of objects.

Object motion The displacement of the center point of the object box between adjacent frames reflects the velocity of the object motion (OM). OM is formulated as $\text{OM}(\mathbf{b}^t, \mathbf{b}^{t-1}) = |x_c^t - x_c^{t-1}| + |y_c^t - y_c^{t-1}|$. x_c and y_c are the horizontal and vertical coordinates of the center of the bounding box, respectively. In the last line of Fig. 8, the interval of OM is 20 pixels. In the BFT dataset, the overall OM is higher compared to other datasets, indicating a higher level of dynamicity within BFT.

8.3. Visualization

The visualization of representative videos contained in the BFT dataset is shown in Fig. 9, which covers all orders of birds in the dataset to demonstrate the diversity of data and the dynamicity of tracking objects.

9. Experiment Details

9.1. Dataset statistics

TAO There are 295 overlapping classes between the 833 object classes in TAO [8] validation set and LVIS [20]. Out of these, 35 classes are considered rare and are designated as *novel* classes following OVTrack [34]. For evaluation purposes, there are a total of 109,963 annotations across 988 validation sequences, with 2,835 annotations belonging to *novel* classes.

TAO-OW TAO-OW [42] validation set considers 52 COCO [36] classes as *known*, which consist of 87,358 distinct object tracks. In contrast, 209 classes that are not present in COCO are regarded as *unknown*, comprising 20,522 distinct object tracks.

AnimalTrack and GMOT-40 Following open-world settings [42], the *known* and *unknown* classes of AnimalTrack [69] test set and GMOT-40 [1] are divided. In the AnimalTrack benchmark, the *known* classes are *horse* and *zebra*, and the *unknown* classes are *chicken*, *deer*, *pig*, *goose*, *duck*, *dolphin*, *rabbit*, and *penguin*. In the GMOT-40 benchmark, the *known* classes are *airplane*, *bird*, *boat*, *sheep*, *car*, and *person*, and the unknown classes are *helicopter*, *billiard*, *lantern*, *tennis*, *balloon*, *fish*, *bee*, *duck*, *penguin*, *goat*, *wolf*, and *ant*.

9.2. Metric details

OWTA The *open-world tracking accuracy* (OWTA) [42] consists of the *detection recall* (DetRe) and *association accuracy* (AssA). With a localization threshold α, the OWTA metric is calculated as $\text{OWTA}_\alpha = \sqrt{\text{DetRe}_\alpha \cdot \text{AssA}_\alpha}$. Specifically, DetRe is evaluated as $\text{DetRe}_\alpha = \frac{\text{TP}_\alpha}{\text{TP}_\alpha + \text{FN}_\alpha}$, where the true positive (TP) and false negative (FP) are considered while false positives (FP) are not penalized. In AssA, the TP associations (TPA), FP associations (FPA), and FN associations (FNA) are incorporated into the calculation as $\text{AssA}_\alpha = \frac{1}{\text{TP}_\alpha} \sum_{c \in \text{TP}_\alpha} \mathcal{A}(c)$, where $\mathcal{A}(c)$ is calucated as $\mathcal{A}(c) = \frac{\text{TPA}_\alpha(c)}{\text{TPA}_\alpha(c) + \text{FPA}_\alpha(c) + \text{FNA}_\alpha(c)}$.

TETA The *tracking-every-thing accuracy* (TETA) [33] consists of three components, the *localization accuracy* (LocA), *classification accuracy* (ClsA), and *association accuracy* (AssocA). TETA is calculated as $\text{TETA} = \frac{\text{LocA} + \text{AssocA} + \text{ClsA}}{3}$. LocA is evaluated as $\text{LocA} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$, and AssocA is derived in the same manner as AssA in OWTA. For classification, ClsA is calculated as ClsA $=$

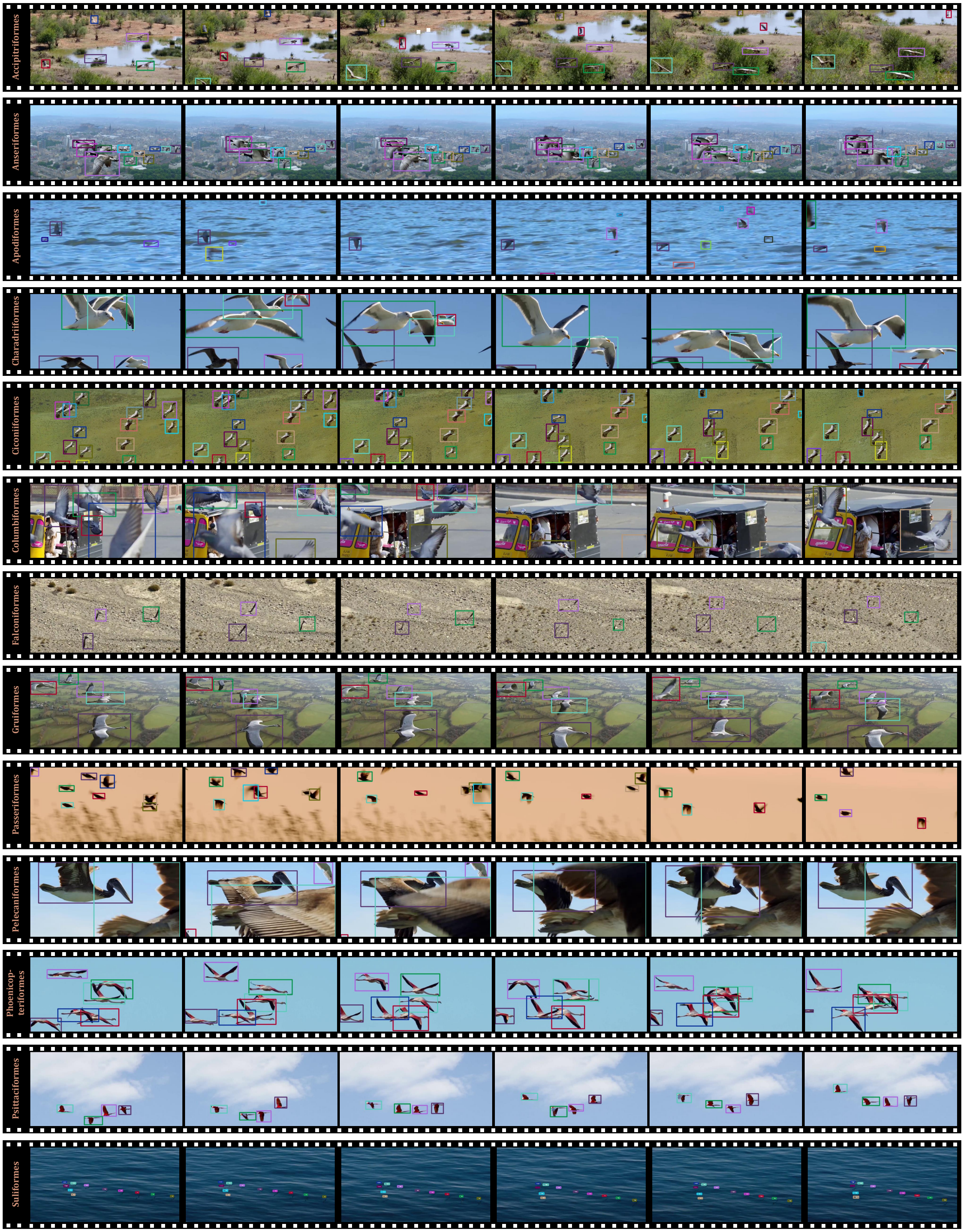

Figure 9. Order-wise visualization of the BFT dataset, showcasing the diversity of the dataset and the dynamicity of tracking objects.

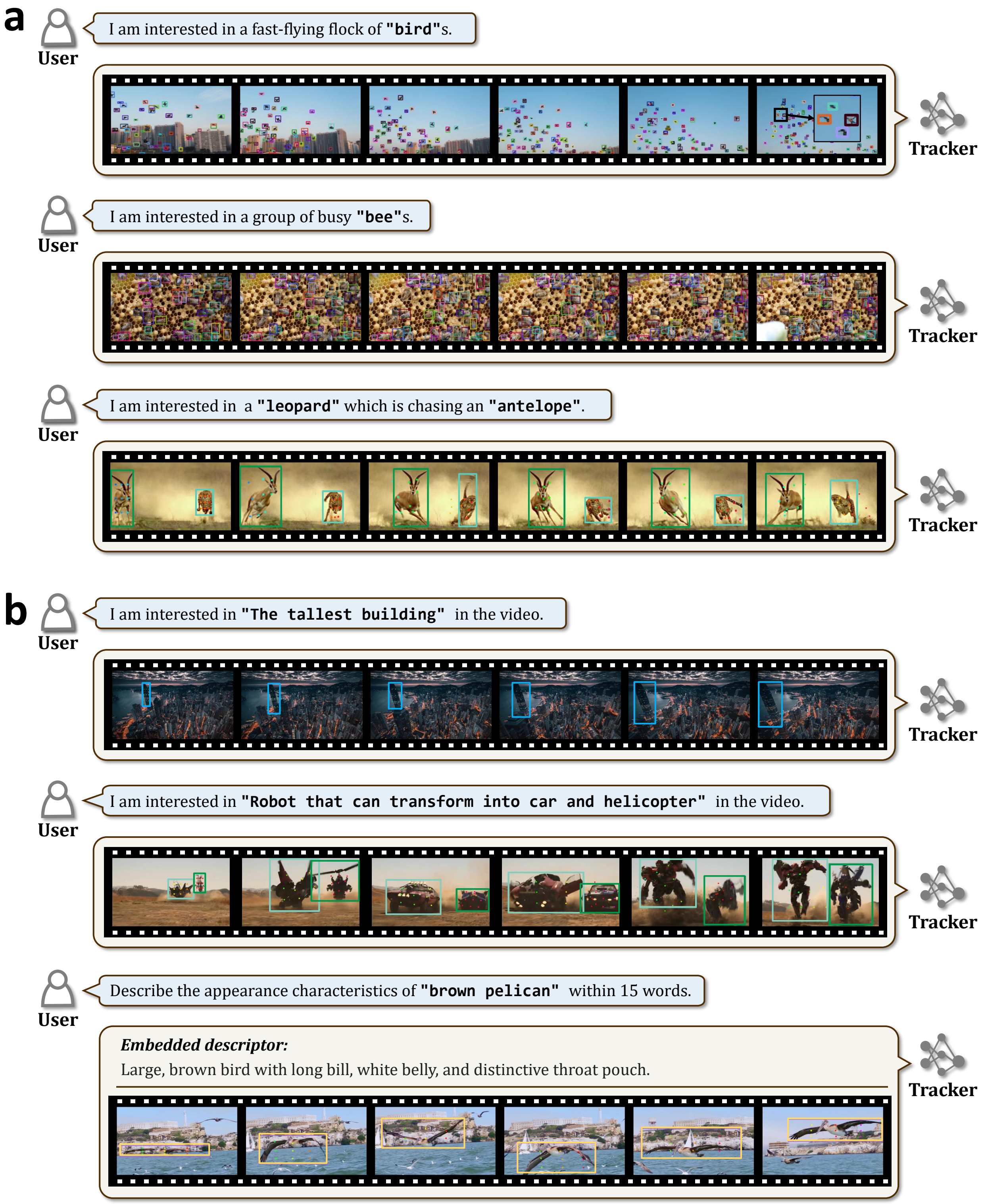
a
User
I am interested in a fast-flying flock of "bird"s.
Tracker
User
I am interested in a group of busy "bee"s.
Tracker
User
I am interested in a "leopard" which is chasing an "antelope".
Tracker
b
User
I am interested in "The tallest building" in the video.
Tracker
User
I am interested in "Robot that can transform into car and helicopter" in the video.
Tracker
User
Describe the appearance characteristics of "brown pelican" within 15 words.
Embedded descriptor:
Large, brown bird with long bill, white belly, and distinctive throat pouch.
Tracker

Figure 10. Qualitative tracking results of NetTrack. **a** Scenarios involving highly dynamic objects and densely packed objects in open-world settings. **b** Understanding dynamic scenes and objects under referring expression comprehension conditions and domain-specific knowledge aided by embedded descriptors.

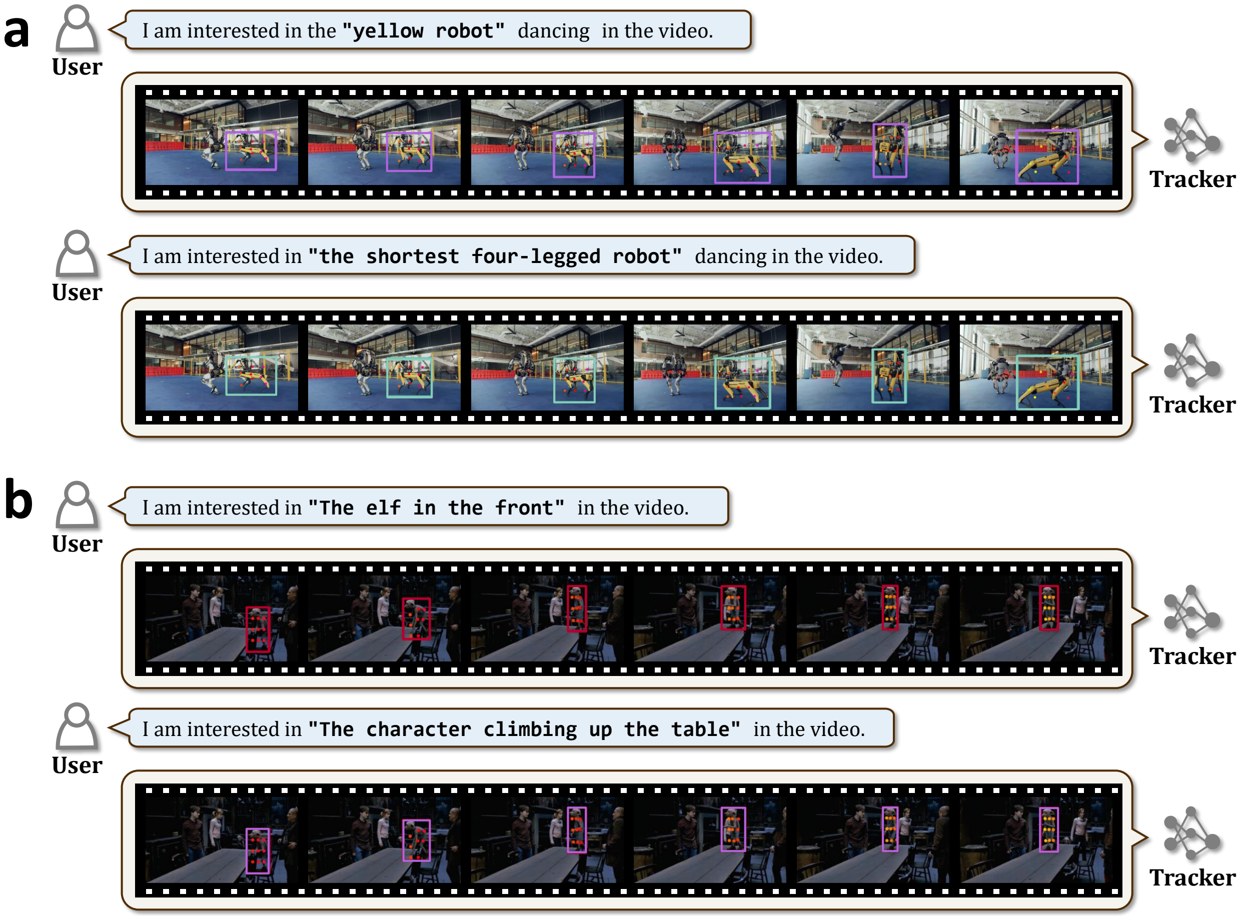
a
User
I am interested in the "yellow robot" dancing in the video.
Tracker
User
I am interested in "the shortest four-legged robot" dancing in the video.
Tracker
b
User
I am interested in "The elf in the front" in the video.
Tracker
User
I am interested in "The character climbing up the table" in the video.
Tracker

Figure 11. Qualitative tracking results to validate the robustness of NetTrack under various prompts. **a** Describing the objects of interest from the perspectives of color and shape. **b** Describing the objects from the perspectives of appearance and behavior.

$\frac{\text{TPC}}{\text{TPC+FPC+FNC}}$, where TP classification (TPC), FP classification (FPC), and FN classification (FNC) are concerned.

HOTA, MOTA, and IDF1 These three metrics are for closed-set MOT and serve as a reference. Higher Order Tracking Accuracy (HOTA) [44] is calculated as $\text{HOTA}_\alpha = \sqrt{\text{DetA}_\alpha \cdot \text{AssA}_\alpha}$. where the detection accuracy (DetA) is derived in the same manner as LocA in TETA. Multiple object tracking accuracy (MOTA) [3] measures the detection errors of FNs and FPs, as well as the association error of identification switch (IDSW), which is derived as $\text{MOTA} = 1 - \frac{\text{FN+FP+IDSW}}{\text{gtDet}}$, where gtDet refers to the number of groundtruth detections. IDF1 [56] is the ratio of correctly identified detections over the average number of ground-truth and computed detections, which balances identification precision and recall through harmonic mean and is calculated as $\text{IDF1} = \frac{2\text{IDTP}}{2\text{IDTP+IDFP+IDFN}}$, where IDTP, IDFP, and IDFN refer to true positive, false positive, and false negative of identification, respectively.

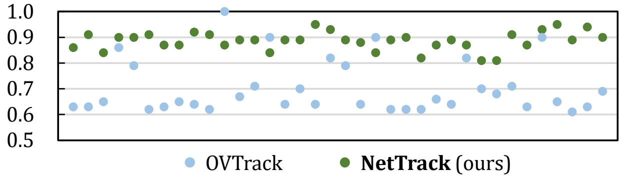

Figure 12. LocA comparison between CLIP-based OVTrack and the proposed NetTrack on BFT benchmark. Each data point refers to a corresponding image sequence.

9.3. Additional ablation studies

Comparison with CLIP-based pre-training Compared to the previous coarse-grained CLIP [54]-based method [34], the proposed fine-grained object-text correspondence demonstrates better capability in localizing dynamic objects. As depicted in Fig. 12, each data point represents the performance of LocA in a sequence from the BFT dataset. Due to the excessive introduction of false positive samples by the compared method, the *association accuracy* on dynamic targets is low (16.5) and thus not included in the com-

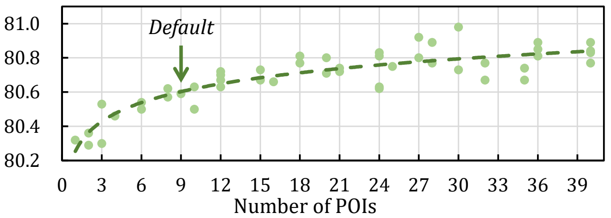

Figure 13. D. Re. performance with different numbers of POIs. More POIs typically bring better performance and heavier computational burden. NetTrack aims to realize a trade-off between performance and efficiency.

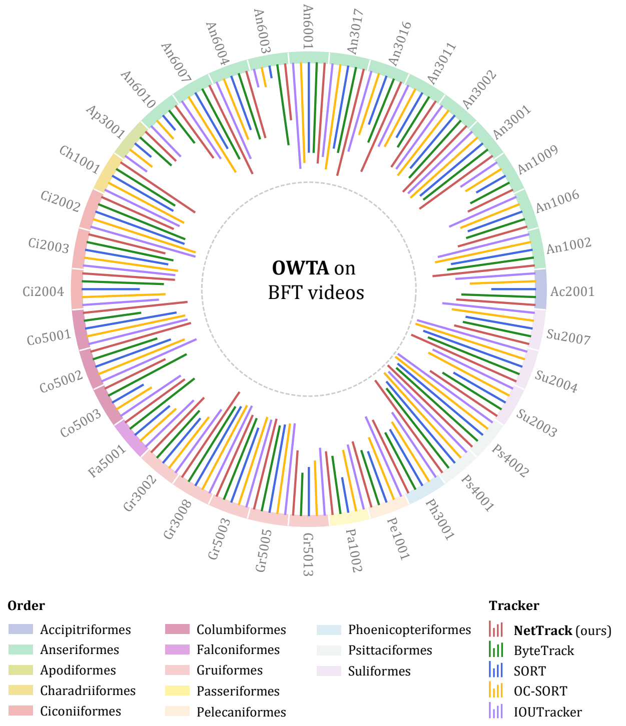

Figure 14. Order-wise performance comparison. NetTrack exhibits stronger generalization ability on diverse categories.

prehensive comparison. This comparison validates the applicability of the introduced fine-grained object-text correspondence to track highly dynamic objects.

Number of POIs Assigning excessive POIs to each potential object can lead to more robust performance, particularly in the ability to discover potential objects. However, it also results in a higher computational burden. Fig. 13 illustrates the relationship between *detection recall* and the number of POIs assigned to each object. Each specific POI count is decomposed into grids corresponding to different aspect ratios, *e.g.*, 12 POIs can be decomposed into 3×4, 4×3, 2×6, and 6×2 grids, resulting in 4 data points. The aspect ratio of the grid is constrained within the range of $[\frac{1}{3}, 3]$. When the number of POIs exceeds 9, the performance improvement becomes marginal. Therefore, the default number of POIs

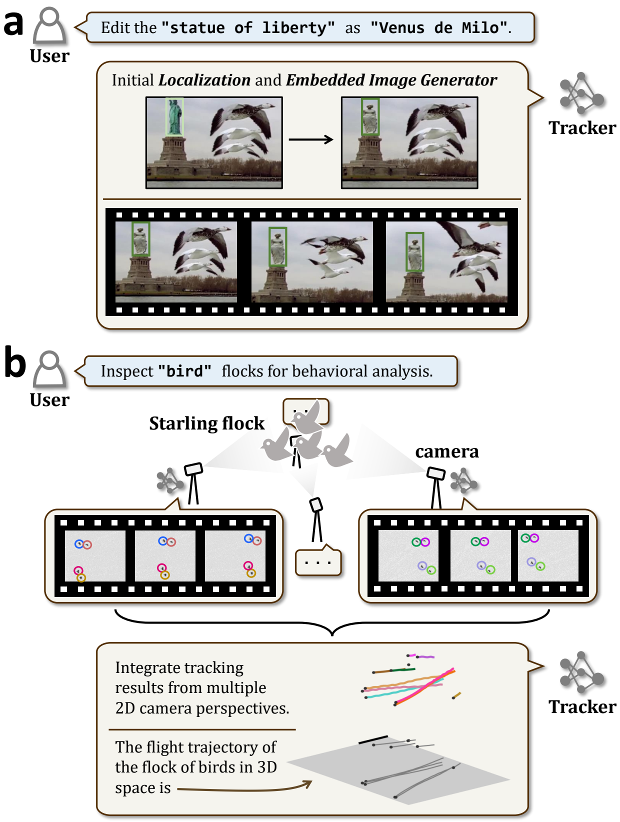

Figure 15. Potential applications of NetTrack. **a** Video editing with embedded image generator, *e.g.*, stable diffusion [57]. **b** Ecological inspection for behavioral analysis, *e.g.*, obtaining 3D trajectories of bird flocks through multi-view tracking results and 3D assignments with settings from [37].

is set to 9, corresponding to a 3×3 grid.

Order-wise performance Fig. 14 categorizes the test videos into the 13 orders included in the dataset and investigates the performance of the tracker on each order separately. NetTrack outperforms other SOTA trackers [4, 5, 7, 71] on most orders, and the performance gap is not significant for different orders, which confirms its strong generalization ability.

9.4. Qualitative results

Tracking dynamic and dense objects Tracking dynamic and numerous objects in an open-world environment poses significant challenges. Fig. 10-a illustrates NetTrack's performance in tracking dynamic and dense objects across three scenarios, including fast-moving and numerous flocks of birds and bees, as well as highly deformable objects like leopards and antelope. Despite these challenges, NetTrack demonstrates excellent robustness.

Referring expression comprehension The understanding of dynamic scenes in open-world environments is crucial for the practical application of trackers in MOT. Fig. 10-b illustrates three scenarios: continuously understanding and tracking the tallest buildings, tracking constantly deforming Transformers, and learning professional knowledge through an embedded descriptor [6]. NetTrack demonstrates its ability to comprehend dynamic scenes and its potential value in practical applications.

Robustness to various prompts In practical applications, the user's prompt input may be biased due to different focuses, making it important for the tracker to have robust performance with diverse prompts. Fig. 11 demonstrates scenarios where the object of interest remains the same, but two different prompts are given. In scenario **a**, prompts focus on the color and height of the robot, respectively, while in scenario **b**, prompts focus on the category of the character and the ongoing action. Faced with diverse prompts, NetTrack is able to maintain robust performance, confirming its potential in practical applications.

10. Applications

In addition to the use of embedded descriptors (*e.g.*, large language models [6, 52]) to understand domain-specific knowledge as shown in Fig. 10-b, Fig. 15 also demonstrates two other potential applications of NetTrack. In Fig. 15-a, an embedded image generator performs inpainting on the objects of interest for video editing. The tracker first locates the objects of interest and then uses the image generator (stable diffusion [57] in this example) to inpaint on the area of interest, achieving the desired video editing effect. Furthermore, in Fig. 15-b, due to NetTrack's use of fine-grained representations of objects, even small and dynamic objects can be tracked by point tracking, enabling the acquisition of three-dimensional (3D) trajectories for ecological inspection, such as bird flight trajectories through multi-view tracking results and 3D assignment. The data and experimental settings for ecological inspection in Fig. 15 are sourced from [37]. It is worth mentioning that these are just brief descriptions of some potential applications of NetTrack. Due to its strong generalization ability, better integration with foundation models will lead to even broader and more practical application value.