Slope-Track: Multiple Object Tracking on Ski Slopes

M'saydez Campbell, Christophe Ducottet, Damien Muselet, Rémi Emonet

Computer Vision and Image Understanding 264 (2026) 104663

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Slope-Track: Multiple Object Tracking on Ski Slopes☆

M'Saydez Campbell [a,*], Christophe Ducottet [a], Damien Muselet [c], Rémi Emonet [a,b]

[a] Université Jean Monnet Saint-Étienne, CNRS, Institut d'Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Étienne, France
[b] Inria; Institut Universitaire de France, France
[c] L3i, La Rochelle Université, F-17000, La Rochelle, France

ARTICLE INFO

Keywords:
Slope-track
Multiple object tracking (MOT)
Analysis
Video monitoring
Winter sports

ABSTRACT

In this paper, we introduce Slope-Track. Slope-Track is a novel multiple object tracking (MOT) dataset designed to reflect the complexities of real ski slope environments. The dataset has over 96,000 frames collected from 10 different ski resorts under various weather and visibility conditions. Slope-Track addresses significant challenges in slope monitoring, including small object sizes, occlusions, fast and irregular motion, and low appearance consistency. It is densely annotated with bounding boxes and object identities, facilitating the evaluation of detection and tracking algorithms. We analyze the dataset's characteristics comparing it to the existing MOT datasets. The results demonstrate that Slope-Track encapsulates a combination of challenges found in other datasets. Additionally, we benchmark a range of existing tracking algorithms and propose a new module that improves motion-based association by dealing with the specific shape of trajectories along ski slopes. Our results demonstrate that incorporating appearance features can have a mixed impact, depending on how they are used within each tracking algorithm. In contrast, motion-based methods and spatial association strategies show more reliable performance. Overall, we provide a challenging benchmark for evaluating and improving multi-object tracking systems in real-world outdoor environments. The dataset and code can be found at https://slopetrack.github.io/.

1. Introduction

Skiing and snowboarding are extremely popular winter recreational sports taking millions of people to ski resorts each year (Association, 2024). The observation of ski resorts, or more specifically of ski slopes, is important to ensure the safety of the persons present on the slopes, the appropriate deployment and the good functioning of specific equipment in the resort, the efficient usage of the resources of the resort and the monitoring of the slope and weather conditions by the visitors. Consequently, this has boosted the use of devices, such as video cameras, to monitor the slopes and their users (Magazine, 2023a,b; Magnusson et al., 2020). This has created a stream of video footage that can be used for a variety of computer vision applications such as detection and tracking and by extension high level understanding tasks such as skier performance level, group detection, path prediction and person time on slope.

Despite the potential use cases of the video data, slope-based activities such as skiing and snowboarding are under-researched compared to other activities such as football (Cui et al., 2023; Cioppa et al., 2022), basketball (Cui et al., 2023) and dance (Sun et al., 2022). This can be attributed to the lack of available annotated data. Most of the available data for skiing and/or snowboarding focuses on pose estimation in 2D and 3D (Bachmann et al., 2019), fall detection (Zwölfer et al., 2023) or single object tracking (Dunnhofer et al., 2024). While these datasets (Dunnhofer et al., 2024; Bachmann et al., 2019; Zwölfer et al., 2021) have some level of object localization, the data is limited, barely annotated or focuses on a single individual. This poses a significant challenge in the development and the applicability of computer vision solutions for slope-based activities.

Video footage of the ski slopes presents unique problems such as adverse weather conditions affecting the field of visibility, varying camera specifications affecting image quality and diverse positions of the camera affecting the point of view. Furthermore, the persons can have fast motion, varying speed, non-linear movements, irregular poses, less recognizable appearance and small size. Understanding the impact of these factors on image-based algorithms is critical for advancing winter-sports vision.

To assist in the advancement of this field, we introduce a new multiple object tracking dataset, incorporating all of the aforementioned

☆ This article is part of a Special issue entitled: 'CV for Sports' published in Computer Vision and Image Understanding.

* Corresponding author.
E-mail addresses: m.saydez.campbell@univ-st-etienne.fr (M. Campbell), ducottet@univ-st-etienne.fr (C. Ducottet), damien.muselet@univ-lr.fr (D. Muselet), remi.emonet@univ-st-etienne.fr (R. Emonet).

https://doi.org/10.1016/j.cviu.2026.104663
Received 30 September 2025; Received in revised form 9 January 2026; Accepted 15 January 2026
Available online 21 January 2026

challenges, called Slope-Track. Slope-Track includes 96,498 frames across 10 different ski resorts, densely annotated with bounding boxes and identities for persons of all ages on the ski slopes regardless of their specific activity. Each video is labeled with the weather conditions and assigned a global level of person occlusion and a level of visibility. Each video is acquired from a single camera which maintains the same position for the duration of this video.

In evaluating existing tracking algorithms on our dataset, we noticed that most algorithms struggle in re-identifying objects that are occluded for more than 15 frames. The utilization of appearance features for re-identification that seem to work with other datasets no longer apply here due to the variation in object features across frames and the utilization of linear motion models such as the Kalman Filter fail due to the irregularity of skier motion patterns. Therefore, we incorporated trajectory modeling for the handling of occluded objects, introducing a new module called GlideTrack.

The main contributions of this paper are:

- The introduction of a new challenging dataset, Slope-Track, covering 10 ski resorts, with diverse weather conditions, where persons have fast motion, varying speed, non-linear movements, irregular poses, less distinguishable appearance and small sizes.
- The benchmarking of the most recent and highly regarded multiple object tracking methods on ski slope-based activities.
- The introduction of a new tracking strategy designed to improve motion-based association by tackling the occlusion problem in ski slope tracking.

In the upcoming sections of the paper, we motivate our work and present our contributions in five main parts. Section 2 reviews prior multiple-object-tracking (MOT) datasets and methods, small-object tracking methods, and computer-vision research for skiing and snow-based activities, highlighting the gaps that motivate our work. Section 3 presents the design, collection, annotation, and statistical analysis of the proposed *Slope-Track* dataset. Section 4 introduces our trajectory forecasting model, *GlideTrack*, describing the Mamba-based motion prediction module (Gu and Dao, 2023; Dao and Gu, 2024) and its integration into a tracking-by-detection pipeline. Section 5.1 summarizes the evaluation metrics used to assess detection and association accuracy, including HOTA (Luiten et al., 2020), MOTA (Bernardin and Stiefelhagen, 2008) and IDF1(Ristani et al., 2016). Section 5 reports benchmark results of state-of-the-art trackers on Slope-Track by analyzing its performance and evaluates the effectiveness of the proposed module. Finally, Section 6 concludes the paper with key findings, discusses limitations, and outlines directions for future work.

2. Related works

2.1. Multiple object tracking datasets

The preexisting multiple object tracking datasets cover a variety of targets, perspectives and objectives such as autonomous driving (Geiger et al., 2012), pedestrian tracking (Milan et al., 2016; Dendorfer et al., 2020), dancers and athletes displaying complex motion (Sun et al., 2022; Cui et al., 2023; Cioppa et al., 2022). For pedestrian tracking, the largest and most used database is the MOT series (MOT17 (Milan et al., 2016) and MOT20 (Dendorfer et al., 2020) datasets). These datasets show pedestrians filmed from different viewpoints in high-density areas such as train stations, intersections, and commercial centers. They are challenging because they require the algorithms to tackle many object occlusions. On the other hand, the objects have regular motion with stable and discriminative appearance over time. Due to this, tracking algorithms that have been developed based on these datasets do not perform well on situations with complex motion. In the dance context, the DanceTrack dataset (Sun et al., 2022) has been proposed to provide situations with limited distinguishable features between objects and irregular motion. This dataset (Sun et al., 2022) concentrates on dancers (large groups or individuals) in motion in different settings such as outdoor, indoor, distant camera and dance in sporting scenes. It forces the construction of tracking solutions that do not depend heavily on the appearance but on the motion of the object. Other datasets requiring similar motion-based solutions are SoccerNet (Cioppa et al., 2022) and SportsMoT (Cui et al., 2023), both focusing on sporting scenes. SoccerNet focuses on football only and annotates almost all the moving elements on the field, while SportsMOT focuses on football, basketball and volleyball and only annotates the players. These datasets highlight motion that is fast with variable speed and similar but unique appearance. Compared to other datasets, our proposed Slope-Track dataset, like DanceTrack, has object instances that have few or no discriminative features for frame-to-frame association and, like SportsMOT, only annotates skiers on the slope, emphasizing fast and variable speed.

2.2. Small object tracking datasets

According to Chen et al. (2022), small objects are objects with sizes less than 32×32 pixels. They naturally occur due to perspective or distance between the object and the camera. There has been increasing interest in research related to issues raised by this situation, and thus, various datasets have been proposed for small object detection (Zhu et al., 2021), tracking (Zhu et al., 2021) and segmentation (Waqas Zamir et al., 2019; Ding et al., 2022). While small objects dominate our dataset, covering approximately 70% of object instances, it is not solely considered a small object dataset. Within the videos, some objects move towards the camera causing them to become larger over time, as others can enter the frame directly close to the camera. Solutions are expected to perform well on both small and large objects.

2.3. Applications of computer vision to skiing/ snow based activity datasets

In recent years, snow activities such as skiing and snowboarding have become an important research field, enabling further analysis of the participants. For example, Zwölfer et al. (2023) introduced a dataset to recognize when alpine skiers are off-balance or falling as additional cues into pose detection. Bachmann et al. (2019) developed a pose estimation dataset for alpine skiing with the purpose of accurately estimating the 3D poses from images, while Ludwig et al. (2023) created a dataset of videos showing ski jumpers for the detection of desired keypoints of limbs and skis. These methodologies include object detection and a degree of tracking for locating the skier position however, it is not completely analyzed.

Recently, a new dataset was established delving into localization and association (Dunnhofer et al., 2024). SkiTb dataset provides high quality single and multi-camera videos of professional skiers from different types of skiing events. It centers on a single skier highlighting different ski slopes, skiing styles and difficult weather conditions. Similarly, our dataset features difficult weather conditions, different ski slopes and skiing styles. However, our dataset provides varying quality (12–30 fps) single camera videos from ski resorts, capturing multiple snow related activities of persons over multiple ski slopes with different proficiency levels. A comparison between SkiTB and the proposed Slope-Track is provided in Table 1.

2.4. Multiple object tracking algorithms

Within the paradigm of multiple object tracking, methods differ primarily in how they model and exploit motion to associate objects across frames. Although most trackers incorporate appearance or spatial similarity through a cost or similarity matrix, their treatment of motion remains the key difference. For example, SORT (Bewley et al., 2016) anchors its associations in motion by using a Kalman Filter to predict object trajectories and compares those predictions to new detections

Table 1
SkiTB is the closest publicly available dataset in the skiing domain. This table highlights key differences between SkiTB and the Slope-Track multi-object tracking dataset, which targets ski-slope scenarios with varied frame rates.

Attribute	SkiTB (Dunnhofer et al., 2024)	Slope-Track
Application	Single object tracking	Multi-object tracking
Activity	Skiing	Skiing, Walking etc.
Weather Annotations	✓	✓
Per-frame Annotations	✓	✓
Complete Trajectory	✓	✓(as best as possible)
Frame Rate	30 fps	12–30 fps (varied)
Avg. Annotated Objects per Frame	1	11.38
Annotated Persons/Athletes ../	196	888
Total Frames	352,978	96,498
Locations	161	10

via IoU. ByteTrack (Zhang et al., 2022) extends this idea by admitting low-confidence detections. BoTSORT (Aharon et al., 2022) further refines motion modeling through camera-movement compensation and an adjusted Kalman Filter. UCMCTrack (Yi et al., 2024) utilizes camera compensation by estimating parameters from a single frame to minimize cost. Additionally, it replaces IoU with a mapped Mahalanobis distance to emphasize predicted motion uncertainty. GeneralTrack (Qin et al., 2024) captures point-wise relations and propagates them to instance-level associations to balance motion and appearance without manual weighting. OC-SORT (Cao et al., 2023) augments the Kalman Filter with object height and velocity for richer motion state estimation. While, Deep-OC-SORT (Maggiolino et al., 2023) adds adaptive appearance features on top of that motion core. HybridSORT (Yang et al., 2024) introduces tracklet-confidence modeling and height-modulated IoU to strengthen the motion-driven discrimination of closely packed objects.

Recent work has focused on other methods that extend past the linear assumptions of the Kalman Filter. Deep-EIoU (Huang et al., 2024) introduces the use of expanding bounding boxes to better capture associations under abrupt motion or partial overlap changes. MeMoTr (Gao and Wang, 2023) adopts a memory-augmented transformer that encodes long-range temporal dependencies to allow the tracker to reason over extended motion histories for robust association through occlusion and abrupt motion changes. CoMOT (Yan et al., 2023) introduces cooperative motion transformers to jointly estimate object dynamics and inter-object motion correlations to better handle crowded or highly interactive scenes.

Other recent methods have increasingly adopted structured state-space models (SSMs) for multi-object tracking (MOT) to replace traditional motion models with more expressive and learnable representations. TrackSSM (Hu et al., 2024) (also referred to as ByteSSM when integrated with Byte) introduces a flow-guided SSM that refines trajectory predictions step-by-step across multiple layers. It encodes each object's historical trajectory using a Mamba model and decodes using a flow-like structure that refines the predicted position at each layer. Similarly, MambaMOT (Huang et al., 2025) employs the original Mamba block coupled with an MLP to predict the next position given an object's history. MambaMOT+ extends this framework with a trajectory embedding head that extracts motion features and optimizes them for use in comparing object trajectories to merge fragmented tracks. Another method that uses Mamba is MambaTrack. MambaTrack (Xiao et al., 2024) uses a bi-Mamba encoder to capture bidirectional temporal dependencies and applies an autoregressive refinement strategy to recover tracklets lost under occlusions. Building on these advances, our approach focuses on enhancing the input representation so that the Mamba encoder can better learn motion patterns to determine the next positions and incorporate it as a dedicated occlusion module explicitly designed for the unique challenges of ski-slope environments.

3. The Slope-Track dataset

This section details how the dataset was constructed and additional information on the properties of the dataset. The Slope-Track dataset is designed for multiple object tracking (MOT) in ski-slope environments. Although its imagery is firmly rooted in the skiing or snow based activities domain, the task shares the same core challenges as established MOT benchmarks such as MOT17/20, DanceTrack, and SportsMOT. Like those datasets, Slope-Track must cope with crowded scenes, frequent occlusions and variability in appearance. The distinguishable factor between the Slope-Track and the other MOT datasets is its unique setting: downhill skiers and snowboarders captured across varied slopes, lighting conditions, and snow states, producing fast yet sometimes irregular trajectories. Therefore, we highlighted how these properties connect Slope-Track to the broader MoT literature and expose new challenges specific to slope-based winter sports.

3.1. Dataset construction

3.1.1. Design

The Slope-Track dataset was designed for the purpose of monitoring individuals on ski slopes for high level understanding tasks. As such, it was important to include the video footage of the persons' full trajectory from the top to the bottom of the slope. It was developed using real videos from ski resorts to ensure that the algorithms created will work in the wild. The dataset includes a range of scenarios, such as varying weather conditions, different viewpoints, low and high quality videos, multiple persons doing varying activities and varying levels of occlusion. This is to ensure that solutions developed can generalize well.

3.1.2. Video collection

Videos of ski slopes were collected from different ski resorts on the internet, each with a duration between 2 min to 6 h. Among these videos, we selected 20 videos between 20 s to 3 min. These videos have been selected for their diversities and complexities. Additionally, we have included 12 videos without annotation but with detected bounding boxes in case of future specific needs.

3.1.3. Video annotation

The annotation process was done by utilizing existing detection and tracking methods to provide initial tracking. Then, the incorrect tracks and very incorrect bounding boxes were corrected manually using the CVAT tool (Corporation, 2024). The initial tracking was established using the Slicing Aided Hyper Inference (SAHI) (Akyon et al., 2022, 2021) method and the DETR model (Carion et al., 2020) to determine the bounded boxes and the Deep-EIoU method (Huang et al., 2024) for association. Only the persons on the ski slopes were annotated, while ignoring the persons on the ski lifts or waiting to get on the ski lifts. For each annotated frame, the annotation labels for each person visible in that frame are: a bounding box label (x,y,w,h format)

Fig. 1. Examples of annotated frames from the Slope-Track dataset illustrating a bounding box in (x,y,w,h) format and its associated numerical identity.

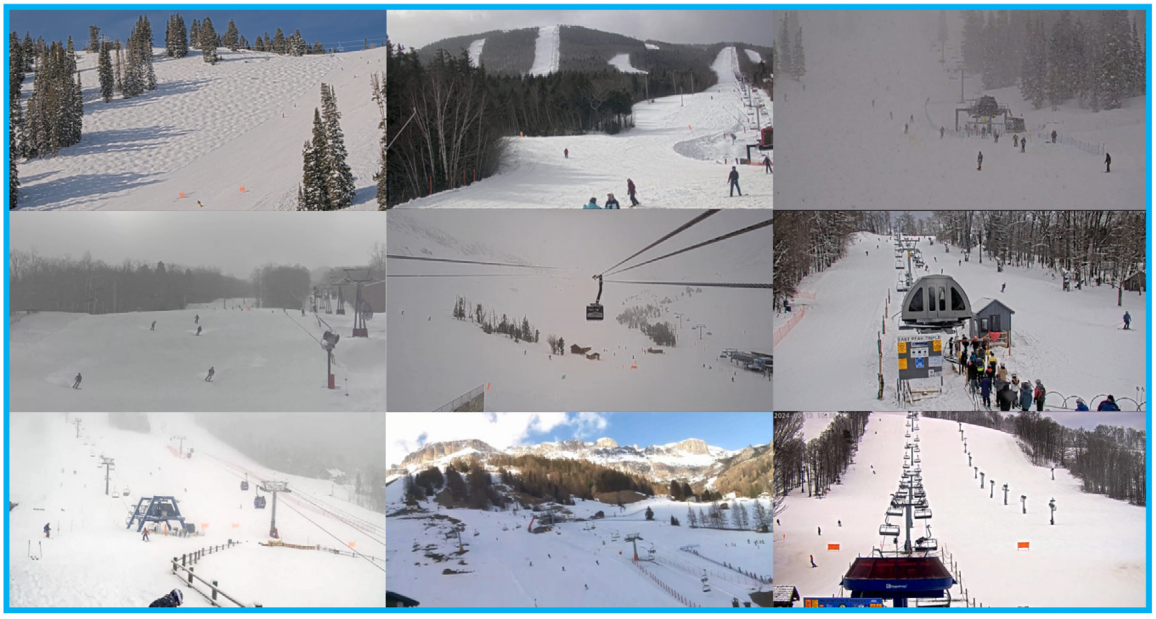


Fig. 2. Examples from Slope-Track dataset showing 9 of the 10 ski slopes included in the dataset. Slope-Track includes diverse weather conditions, visibility levels and viewpoints (unique camera position and camera quality for each viewpoint).

and an identity label (numeric identifier), following the MOT datasets annotation format (Milan et al., 2016; Dendorfer et al., 2020) as seen in Fig. 1. Fully occluded persons were not annotated until they reappear in a future frame. For partially occluded persons, only the visible part of the person is annotated. Persons that completely disappear behind other objects (trees, other persons, etc.) in the scene and cannot be re-identified are assigned a new identity.

3.2. Dataset statistics

3.2.1. Dataset split

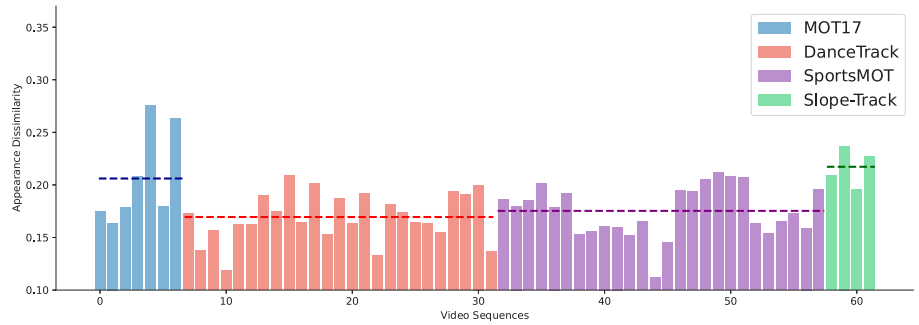
We collected **32** videos of varying lengths between 30 s to 3 min utilizing 11 videos for the training set, 4 videos for the validation set, 5 videos for the test set and an additional set of 12 videos that can be used as an unlabeled test set. The dataset is **3318** seconds in length with **96,498** frames (including the unannotated test set), **555,528** bounding boxes and **888** unique identities (corresponding only to the annotated videos). The splits are prepared such that they offer large inter and intra diversity in viewpoints and weather conditions. However, the dataset will be completely available online and any different future split can be applied depending on the considered application.

3.2.2. Scene diversity

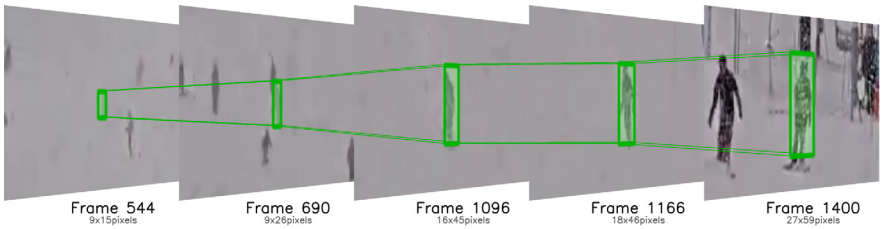
The Slope-Track dataset includes clips from 10 different ski slopes. While each clip is acquired by a single fixed camera, the viewpoints and image quality are different across clips. Obviously, the weather conditions are also varying between videos. Within the dataset, there are four distinct weather conditions used to describe a scene: cloudy, fog, snowing, sunny. The descriptors fog and snowing are split into three levels from the lowest to highest level. These weather conditions can be combined to describe a scene accurately. Some images extracted from our dataset are shown in Fig. 2.

3.2.3. Appearance features

As can be seen in Fig. 3(b), the dataset includes some persons that can be considered under the small object paradigm. Due to this, a



(a) Mean inter-frame cosine distance of re-ID features



(b) Example showing the change in features for the same object in the Slope-Track dataset including bounding box size in pixels.

Fig. 3. (a) Mean inter-frame cosine distance of re-ID features by video identity. Slope-Track has the highest overall mean inter-frame cosine distance in comparison to the other datasets. In other words, the appearance similarity between instances of the same object is lower. The dashed line represents, for each dataset, the average inter-frame cosine distance of re-ID features across all the videos of that datasets. (b) An example showing the change in features for the same object in the Slope-Track dataset, including bounding box size in pixels.

person can have low resolution and background interference which makes it difficult to extract appearance features for association across frames. In order to measure the complexity of the association step based on the appearance features in our dataset, we conducted a study that measures the evolution of features across time (different distances of a person relative to the camera). To compare object appearance similarity between the multiple object datasets, we utilize a pretrained re-identification (re-ID) model (Pei, 2019) to extract the appearance features of objects in each frame and compute the average inter-frame features cosine distance for each instance of a given object along its trajectory and report the mean over all objects of a given video as:

$$\frac{1}{|I|}\sum_{i\in I}\frac{1}{m_i^2}\sum_{k=1}^{m_i}\sum_{l=k+1}^{m_i}\left[1-\cos\left\langle F(S_i^k),\,F(S_i^l)\right\rangle\right] \tag{1}$$

where I a set of unique identities, m_i is the number of frames where an object i appears, $F(S_i^k)$ are the extracted appearance features of object S on frame k from id i and $\langle.\rangle$ is the angle between two vectors.

We noticed that in comparison to the other existing multiple object tracking datasets, Slope-Track has the highest mean inter-frame object dissimilarity as seen in Fig. 3(a). This is due to the variance in the features of the objects across frames.

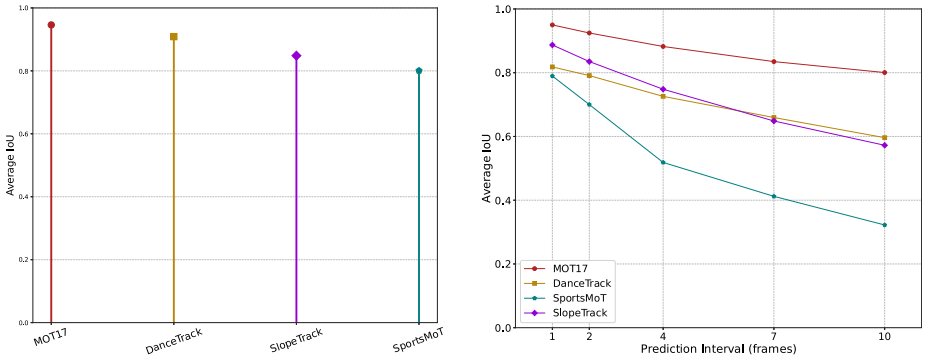
3.2.4. Motion analysis

It is very important to understand motion cues for tracking an object across frames. Each of the pre-existing datasets has a distinct motion pattern which impacts the development of tracking algorithms. The MOT series (MOT17 and MOT20) focuses on pedestrians in crowded scenes, moving at constant speed, with linear motion. DanceTrack concentrates on dance videos with diverse movements but with a low speed. SportsMOT emphasizes fast and variable speed with changing direction. The motion of the Slope-Track dataset is relatively smooth, fast and consistent as persons move on the slope, however they can abruptly increase or decrease their velocity or even stop as they move down the slope.

Utilizing a metric introduced by Sun et al. (2022), it can be seen in Fig. 4(a) that Slope-Track has a score similar to SportsMOT when calculating the IoU (Intersection over Union) measure on corresponding ground truth bounding boxes of adjacent frames, showing that the objects have fast motion compared to their frame rate. Following SportsMoT analysis (Cui et al., 2023), Kalman Filter-based IoU measure was calculated. Instead of focusing on adjacent frames only, the evaluation was done over different prediction intervals in order to reveal the ability of the model to handle complex motion over time. The IoU was calculated between the ground truth and the prediction at the specified interval, for example every 10 frames. As seen in Fig. 4(b), Slope-Track shows good performance when predictions are made frequently, reflecting the smooth and predictable motion of skiers in short time spans. However, the prediction quality deteriorates as the number of frames between predictions increases. This highlights the presence of variable-speed and non-linear motion in the dataset caused by turns, speed changes, and slope variations.

3.3. Limitations

There are three limitations of the proposed dataset. Firstly, there can be instances where the bounding boxes are not exactly positioned around persons on the slope. However, given the nature of our intended application, we prioritize recall over precise detection. As such, we do not expect this to drastically impact the performance of algorithms. Secondly, this dataset is currently smaller in scale compared to some existing multi-object tracking datasets. This limitation can be restrictive for data-hungry models, such as end-to-end trackers, which require extensive and diverse training data to achieve strong generalization. Thirdly, the dataset is captured entirely from a single-camera perspective. While this reflects the intended deployment scenario, it limits the exploration of multi-camera tracking (MTMC) challenges such as cross-camera identity association, viewpoint variation, and coordinated tracking across overlapping fields of view.



(a) IoU on adjacent frames

(b) Kalman Filter-based IoU on different prediction intervals

Fig. 4. (a) IoU on adjacent frames. Slope-Track has fast motion comparable to SportsMOT. (b) Kalman Filter-based IoU on different prediction intervals. Slope-Track has smooth and consistent motion with abrupt changes in velocity and direction.

4. GlideTrack: a new sequence to sequence trajectory forecasting model

In this section, we present our trajectory forecasting model, GlideTrack. It was designed to improve motion-based association by predicting the future position of a lost object and should work in conjunction with other trackers in the tracking by detection paradigm such as DeepEIoU, ByteTrack or Hybrid-SORT.

Firstly, we describe the utilization of the MAMBA (Gu and Dao, 2023; Dao and Gu, 2024) architecture to encode the spatial and temporal information from the previous positions, which is used to predict the future possible positions of an object. Then, we explain how it can be applied to the existing trackers using Deep-EIoU as an example. Lastly, we present details for training and inference.

4.1. Problem formulation

We address the task of multi-object trajectory forecasting on continuous ski-slope videos captured by fixed cameras. Let the video be a sequence of frames

$$\mathcal{V} = \{F_0, F_1, \dots, F_T\}, \tag{2}$$

where T is the total number of observed frames. At each frame t, every visible participant i (skier, snowboarder, or pedestrian) is represented by a bounding box

$$\mathbf{b}_t^i = (c_t^{x,i}, c_t^{y,i}, w_t^i, h_t^i) \in \mathbb{R}^4, \tag{3}$$

with center $(c_t^{x,i}, c_t^{y,i})$, width w_t^i, and height h_t^i. Initial detections will come from an automated person detector.

Goal. Given the history of B objects observed up to frame t,

$$\mathcal{H}_t = \{\mathbf{b}_0^{1:B}, \dots, \mathbf{b}_t^{1:B}\}, \tag{4}$$

our objective is to predict the next K bounding boxes for objects who have been lost due to occlusion or missed detections.

4.2. GlideTrack

We represent each object's motion as a sequence of normalized bounding boxes. Each bounding box is first transformed into a spatial embedding by passing its normalized coordinates through a bank of sine–cosine frequency bands. The resulting spatial representation is then combined with the normalized bounding-box values themselves. This combined sequence is linearly projected and encoded by a single Bi-Mamba model following (Xiao et al., 2024): one pass processes the sequence forward in time, while a second pass operates backward, and their hidden states are concatenated to capture long-range temporal dependencies. Two MLP decoders then read the encoded features to predict the next (K) bounding boxes. The complete architecture is illustrated in Fig. 5. Training employs a weighted combination of Smooth L1 and GIoU losses to balance precise localization with high spatial overlap, enabling accurate multi-step trajectory forecasting.

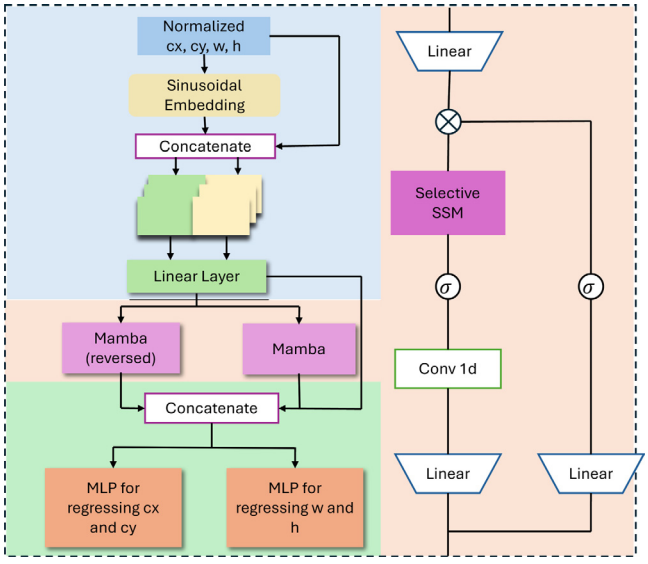


Fig. 5. GlideTrack architecture: past normalized bounding boxes are encoded into spatial and content embeddings, fused into a unified representation and processed by a sequence-to-sequence Mamba motion predictor.

Input embeddings. Using the normalized bounding boxes, we extracted two complementary embedding to encode different aspects of the data. Each raw box sequence $\{(c_x, c_y, w, h)_1, \dots, (c_x, c_y, w, h)_t\}$ is first normalized relative to its starting frame and the average box size over the t-frame context. For each bounding box at time step t, we compute;

$$\tilde{c}_x = \frac{c_x^{(t)} - c_x^{(1)}}{\frac{1}{t}\sum_{i=1}^{t} w_i}, \qquad \tilde{c}_y = \frac{c_y^{(t)} - c_y^{(1)}}{\frac{1}{t}\sum_{i=1}^{t} h_i}, \tag{5}$$

$$\tilde{h} = \frac{h^{(t)} - h^{(1)}}{\frac{1}{t}\sum_{i=1}^{t} h_i}, \qquad \tilde{w} = \frac{w^{(t)} - w^{(1)}}{\frac{1}{t}\sum_{i=1}^{t} w_i}. \tag{6}$$

This anchors all trajectories to their first-frame position and scales them by the average width and height of the observed sequence, providing invariance to absolute location and object scale. The choice of normalization will be discussed further in Section 5.4.2.

We encoded the spatial positioning information using a similar formulation introduced in Vaswani et al. (2017). We use this formulation because it provides a smooth and scale-free way to represent absolute coordinates while allowing the model to infer relative distances through simple linear operations (Yan et al., 2021). Therefore, we expect the representation to highlight the shape and dynamics of the trajectories when applied to normalized boxes. Each component of the bounding box vector $\mathbf{b}_t$ is expanded through a bank of sinusoidal frequency bands;

$$\omega_k = 10000^{2k/d}, \qquad k = 0, \dots, \frac{d}{2} - 1. \tag{7}$$

The mapping

$$\gamma(v) = \left[\sin(v\,\omega_0), \cos(v\,\omega_0), \dots, \sin(v\,\omega_{\frac{d}{2}-1}), \cos(v\,\omega_{\frac{d}{2}-1})\right] \tag{8}$$

is applied to the normalized center coordinates c_t^x, c_t^y as well as the width and height w_t, h_t. In our case, d is set to 8. A discussion on the value d can be found in Section 5.4.3. Concatenating these results yields the spatial embedding

$$\mathbf{s}_t \in \mathbb{R}^{4\times d}, \tag{9}$$

which is then projected to the dimension 32 through a linear layer. Finally, we enrich the representation with the absolute normalized bounding-box values, $\mathbf{c}_t$.

The complete input at each time step is then formed as

$$\mathbf{z}_t = (\text{Linear}(\text{concat}(\mathbf{c}_t, \mathbf{s}_t))). \tag{10}$$

Mamba sequence model. The sequence $\{\mathbf{z}_1, \dots, \mathbf{z}_t\}$ is processed by a single-layer Bi-Mamba network. Each Mamba layer is inherently unidirectional and observes the sequence only from the first to the last step. To overcome this limitation, we also feed the model the sequence in reverse order, allowing it to capture richer patterns and more intricate dependencies across the trajectory.

$$\begin{aligned} \mathbf{h}_t^{\text{fwd}} &= \text{Mamba}_{\text{fwd}}(\mathbf{z}_1, \dots, \mathbf{z}_t), \\ \mathbf{h}_t^{\text{bwd}} &= \text{Mamba}_{\text{bwd}}(\mathbf{z}_t, \dots, \mathbf{z}_1), \\ \mathbf{h}_t &= \text{concat}(\mathbf{h}_t^{\text{fwd}}, \mathbf{h}_t^{\text{bwd}}, (\mathbf{z}_1, \dots, \mathbf{z}_t)), \end{aligned} \tag{11}$$

Trajectory decoding. Two separate multilayer perceptrons (MLPs) are employed to predict the next K bounding boxes, one dedicated to the normalized center coordinates (x, y) and the other to the normalized width–height pair (w, h):

$$\begin{aligned} \hat{\mathbf{c}}_{t+1:t+K} &= \text{MLP}_{xy}(\mathbf{h}_t), \\ \hat{\mathbf{s}}_{t+1:t+K} &= \text{MLP}_{wh}(\mathbf{h}_t). \end{aligned} \tag{12}$$

The final bounding box sequence is obtained by concatenating these outputs,

$$\hat{\mathbf{b}}_{t+1:t+K} = \left[\, \hat{\mathbf{c}}_{t+1:t+K} \;;\; \hat{\mathbf{s}}_{t+1:t+K} \,\right], \tag{13}$$

where both MLPs share parameters across the K steps to generate all future positions.

Training objective. We supervise with a weighted combination of regression and overlap losses. This choice is motivated to ensure that the predicted positions are numerically close to the target and encourage the predicted boxes to have a high spatial overlap which is important for tracking.

$$\mathcal{L} = \lambda_1\, \text{SmoothL1}(\hat{\mathbf{b}}, \mathbf{b}) + \lambda_2\, \text{GIoU}(\hat{\mathbf{b}}, \mathbf{b}), \tag{14}$$

where $\hat{\mathbf{b}}$ and $\mathbf{b}$ denote predicted and ground-truth boxes over all future frames. Weights λ_1, λ_2 balance localization accuracy and intersection-over-union quality.

This design allows Mamba to exploit rich spatial encodings, and content information to produce accurate, multi-step forecasts of object trajectories.

Training. The trajectory forecasting model is trained to predict future object locations from sequences of past observations with a history of 60 positions and a horizon of 60. We applied random scaling and translation data augmentations to each sequence to improve generalization. Training uses a batch size of 128 and the AdamW optimizer with a weight decay of 10^{-4}. The learning rate is warmed up linearly for the first 100 epochs and then decayed using a cosine schedule to 10^{-5} over the remaining 600 epochs, starting from a base learning rate of 10^{-4}. Supervision combines Smooth L1 regression with a beta of 0.001 and GIoU losses each weighted at 0.5. The model is trained for 700 epochs.

4.3. GlideTrack + DeepEIoU

Here, we present how to integrate our sequence-to-sequence module in DeepEIoU (Huang et al., 2024).

The training pipeline includes three main components: the detection model, the ReID model, and GlideTrack. Each element was trained separately so that it could be used in other tracking-by-detection algorithms for fair comparison. For detection, we replaced the existing detector with the YOLOv11 large (YOLOv11l) version (Jocher et al., 2023), enhanced with SAHI (Akyon et al., 2022) to improve small object detection. For appearance-based association, we re-train the ReID model proposed in Yang et al. (2024) to provide embeddings tailored to our dataset. Additional details on model configurations and training settings can be found in Section 5.2.2. GlideTrack is trained as a sequence-to-sequence module to predict object motion as described in Section 4.2.

During inference, the pipeline consists of three steps: detection, motion prediction using GlideTrack, and data association with DeepEIoU (Huang et al., 2024). The detector first provides candidate bounding boxes for each frame. Objects that remain continuously observed are handled using the standard DeepEIoU procedure, where the last positions of active tracks and the incoming detections are spatially expanded and associated using IoU and, optionally, appearance features (ReID). GlideTrack is invoked only when an object is missed. In such cases, the model predicts the object's next position for as long as it remains unobserved, until it is successfully matched or exceeds the maximum allowed missing-frame count. The most recent predicted position of the lost track is then treated as its last known location and passed to DeepEIoU for association.

Trajectory-aware association. For each active track i, we maintain its most recent confirmed sequence of bounding boxes

$$\mathbf{b}^i_{t-L+1:t} = \{(c_x, c_y, w, h)_{t-L+1}, \dots, (c_x, c_y, w, h)_t\}, \tag{15}$$

where L is the observation window. When a track is not observed in the next frame, we invoke the GlideTrack sequence model (Section 4.2) to predict the object's location for every frame it has been missing. If track i was last seen at time t and remains unseen until $t+\delta$, we estimate

$$\hat{\mathbf{b}}^i_{t+1:t+\delta} = \text{GlideTrack}\left(\mathbf{b}^i_{t-L+1:t}\right), \tag{16}$$

providing predicted positions for all intermediate steps $t+1, \dots, t+\delta$. For example, if an object is absent in frame 2 and still missing in frame 3, the model outputs predicted locations for both frame 2 and frame 3.

Detection-to-track matching. Detections are matched to tracks using DeepEIoU (Huang et al., 2024). We match the latest forecast $\hat{\mathbf{b}}^i_{t+\delta}$ for each lost track to the detections at frame $t+\delta$ using DeepEIoU. Due to the unreliability of appearance, we downscale the embedding distance before fusing it with the IoU distance. This ensures that IoU remains the primary driver of association and that appearance features only resolve cases where IoU is ambiguous.

Track management. The missing positions of lost tracks continue to receive updated GlideTrack forecasts until a maximum age is reached or a matching detection is found. Tracks that have been matched consistently across frames continue to do so without prediction. New identities are spawned for unmatched detections.

Bounding box normalization and updates. The trajectory model is trained on normalized bounding boxes. As a result, its outputs must be denormalized to the image coordinate space. For every active track i, a moving average of height and width is maintained and updated whenever a match is made. If track i was last observed at frame t and remains unobserved, the forecast for frame $t+1$ is denormalized using the most recent running average.

Buffer maintenance. Each track maintains a history buffer of at most 60 frames containing its most recent verified detections. The predicted boxes are excluded from this buffer unless a matching detection subsequently confirms them. It is only after this confirmation that the detection for that frame is appended to the buffer. This ensures that the sequence used for future predictions contains only validated bounding boxes, thereby reducing the risk of drift.

Rolling forecast. The sequence-to-sequence architecture of the trajectory model constrains the forecast horizon to the number of confirmed observations in the history buffer. When the buffer for a given track contains L verified detections (for example, L=45), the model can predict at most the subsequent L steps in a single pass. To accommodate longer occlusions, we use GlideTrack in a rolling manner: after an initial L–step forecast is produced, the oldest confirmed observations are discarded, the most recent predictions are used as the new input window, and the model is called again. This process continues until the track is re-detected or reaches its maximum allowed age.

4.4. GlideTrack + ByteTrack

Here, we present how GlideTrack can be integrated into the ByteTrack (Zhang et al., 2022) . Unlike DeepEIoU (Huang et al., 2024), ByteTrack (Zhang et al., 2022) relies exclusively on IoU-based association and a Kalman Filter (KF) for motion modeling. ByteTrack does not use appearance features, therefore, the motion model plays a central role in maintaining track continuity. Integrating GlideTrack into ByteTrack introduces additional considerations, as GlideTrack does not produce the complete 8-dimensional KF state. The training pipeline, track management, bounding box normalization and updates, buffer maintenance, and the rolling forecast mechanisms are the same as described in Section 4.3.

Hybrid motion model. ByteTrack predicts future states using a Kalman Filter before performing IoU-based matching. GlideTrack produces only bounding-box geometry and does not output any additional information. Therefore, we integrated it in a complementary manner. The Kalman Filter continues to propagate both the motion state and its covariance for all active tracks. GlideTrack is invoked only when an object becomes unobserved. In such cases, the Kalman Filter's positional estimate is replaced by the most recent GlideTrack prediction. This preserves the Kalman Filter's uncertainty modeling while allowing the tracker to exploit GlideTrack's non-linear motion predictions during gaps.

Detection-to-track matching. Tracks that remain consistently observed are matched to detections using the standard IoU metric, following the original ByteTrack procedure. When a track becomes lost, its future position is first predicted using GlideTrack. To increase tolerance to localization errors during recovery, both the predicted box and candidate detections are expanded by a factor of $e = 0.5$ before IoU-based matching (Huang et al., 2024).

Possible negative impacts. Although GlideTrack improves non-linear motion handling, its integration into ByteTrack can introduce several drawbacks. GlideTrack does not output velocity or uncertainty estimates therefore, there can be a mismatch between its predictions and the Kalman Filter state. After a track is reacquired, the KF update may become inconsistent with the GlideTrack-generated positions, occasionally producing unstable corrections. As such, the resulting trajectories may not be as smooth.

5. Experimental evaluation

To assess the performance of GlideTrack and evaluate the Slope-Track dataset, we conduct a series of experiments that evaluate the capabilities of multiple object trackers. All experiments are performed using the same detection and re-identification model (where applicable) to ensure fair comparisons across trackers. This allows us to analyze how different motion modeling, association strategies, and the incorporation of appearance features influence tracking performance on the Slope-Track dataset. Finally, we analyze the performance of GlideTrack and its design elements.

5.1. Evaluation metrics

The main metrics that are used to evaluate multiple object tracking are Higher Order Tracking Accuracy (HOTA) (Luiten et al., 2020) and Multiple Object Tracking Accuracy (MOTA) (Bernardin and Stiefelhagen, 2008), with HOTA dethroning MOTA as the best evaluation metric. HOTA equally evaluates detection and association thereby, producing a more accurate evaluation of the result in comparison to MOTA. Additionally, HOTA can be decomposed into detection accuracy (DetA) and association accuracy (AssA) allowing for the separate assessment of different parts of the algorithm. Another metric that can be useful to determine the quality of tracking performance is the identity F1 score (IDF1) (Ristani et al., 2016). The IDF1 metric measures the precision and recall of the tracking algorithm in terms of identity preservation.

5.2. Experimental setup

5.2.1. Dataset configuration

Slope-Track split remains the same as described in Section 3.2.1.

5.2.2. Model configuration

To ensure that the comparison is fair, we replaced the detection model and the re-identification models in the tested trackers with models trained on our dataset, except for the end-to-end methods. We replaced the existing detection model YOLOX (Ge et al., 2021) with the YOLO version 11 large (YOLOv11l) (Jocher et al., 2023) enhanced with SAHI (Akyon et al., 2022). Using the COCO pretrained weights, the YOLO11 large (YOLO11l) model was finetuned utilizing the Slope-Track training set for 100 epochs. We froze the backbone of the model and adjusted the image size (1088), batch size (36), optimizer (AdamW) and learning rate (0.0000001). For pre-processing, we split the frames into 640 × 640 slices with an overlap of 0.1. Additionally, we adjusted the data augmentation strategies; mosaic (0.2), scale (0.8) and removed erasing. For methods that emphasize generalizability, such as GeneralTrack, we retain their original appearance models without modification. For other methods using an appearance model, we retrained the ReID model proposed by Yang et al. (2024), following their default parameters and instructions. Additionally, to show the effectiveness of our method, we extended the duration for which a missing object is retained in memory to 350 frames for every tracking algorithm.

5.3. Benchmark results

We benchmark the Slope-Track dataset with a wide range of modern tracking methods to evaluate how different association strategies cope in slope-based tracking. Table 2 summarizes the results for all the tested methods. These results show that tracking performance varies considerably depending on the underlying motion modeling, association strategy, and use of appearance features.

The table is organized into three main sections: no explicit motion modeling, Kalman filter-based trackers and learned motion models. For each method, we report metrics including HOTA, IDF1, AssA, MOTA, DetA, and LocA. A checkmark indicates the use of appearance features,

Table 2
Benchmarking results of recent tracking-by-detection and learned-motion tracking algorithms on Slope-Track. ✓ indicates the use of appearance features (e.g., ReID).

Algorithm	AF	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑	LocA↑
DeepEIoU		65.0	66.7	58.4	74.2	72.4	94.8
DeepEIoU	✓	67.4	69.0	62.9	74.2	72.3	94.8
Kalman Filter							
ByteTrack		62.7	64.9	56.6	74.9	69.5	90.6
OC-SORT		62.7	61.6	54.7	72.8	71.9	94.4
Deep-OC-SORT	✓	64.3	64.5	60.0	70.5	69.0	**95.5**
HybridSORT		66.6	66.2	62.7	71.0	70.7	94.4
HybridSORT	✓	59.8	59.3	49.2	74.8	72.7	94.9
BoTSORT		65.2	66.9	60.4	73.9	70.3	93.1
BoTSORT	✓	65.8	67.9	61.5	73.9	70.4	93.1
GeneralTrack	✓	59.9	57.1	49.0	75.2	**73.4**	94.7
Learned Motion							
End-to-End							
MeMoTR	✓	61.7	71.2	59.7	**76.1**	64.1	85.8
COMOT	✓	55.7	64.5	58.1	61.4	53.7	86.1
Mamba-based							
ByteSSM		58.7	57.8	50.6	68.2	68.1	95.3
DeepEIoU + GlideTrack		**68.0**	70.3	63.6	74.7	72.8	94.8
DeepEIoU + GlideTrack	✓	**68.4**	**71.1**	**64.4**	74.7	72.6	94.8
ByteTrack + GlideTrack		63.7	68.3	58.8	74.2	69.0	90.6

such as ReID embeddings, which in some trackers stabilize identity tracking but in others may degrade performance depending on the integration strategy.

Among the methods, Deep-EIoU (Huang et al., 2024) is notable despite lacking an explicit motion model. Its strategy of expanding bounding boxes absorbs sudden direction changes and unpredictable skier speeds, helping to preserve identities in fast motion. However, Deep-EIoU still often loses targets when they go undetected for even a few frames. The addition of appearance features can improve its ability to maintain identity consistency across short-term occlusions.

Among the Kalman-based trackers, BoT-SORT (Aharon et al., 2022), HybridSORT (Yang et al., 2024), and OC-SORT (Cao et al., 2023) refine the standard filter with improved state representations and association logic representing the best results in this section. These adjustments provide some benefit but do not completely remove the linear-motion assumption. Therefore, all four methods still drop identities during the fast, erratic skiing and occlusions typical of Slope-Track (Fig. 6). The addition of appearance cues produces mixed outcomes. BoT-SORT gain stability when visual embeddings are included, and Deep-OC-SORT improves over plain OC-SORT, whereas HybridSORT degrades.

Among the learning-based methods, Deep-EIoU combined with GlideTrack demonstrates the best overall performance on Slope-Track. Even without appearance cues, this combination outperforms all Kalman-based and other learned-motion trackers by maintaining trajectory consistency in challenging scenes. By including appearance features, we further improve the results. Figs. 6 illustrate how GlideTrack remains close to the ground truth, even in sequences that cause persistent drift or fragmentation in competing trackers.

We compare GlideTrack with the only publicly available Mamba-based MOT tracker at this time, ByteSSM. ByteSSM struggles on Slope-Track for several reasons. First, it predicts the next position one step at a time rather than forecasting multiple future positions. This can accumulate errors over long sequences. Secondly, it relies on a short historical window of 3–5 past positions to make predictions. This is not the best choice for the Slope-Track dataset because trajectories are relatively consistent and can benefit from a larger history. Third, this short-term focus limits ByteSSM's ability to handle occlusions effectively. The model lacks the temporal context needed to maintain identity continuity through abrupt movement or partial occlusions.

End-to-end trackers such as COMOT and MeMoTR rely heavily on appearance cues and transformer- or memory-based reasoning, which are typically data-hungry and require large amounts of training data to generalize effectively. On Slope-Track, the relatively small dataset size limits their ability to learn robust representations, causing the models to miss detections when targets are small or partially occluded.

5.4. Exploratory analysis

We conduct an exploratory analysis to better understand the factors that influence GlideTrack's performance on Slope-Track. In particular, we investigate how different ski slope viewpoints impact its ability to maintain accurate and consistent trajectories, assess the effectiveness of our proposed module through targeted ablations, and compare several normalization strategies that address the dataset's variability in scale, perspective, and motion dynamics. These analyses highlight both the robustness and the limitations of GlideTrack when applied to diverse and challenging slope-based scenarios.

5.4.1. Analysis of different ski slope viewpoints

As seen in Table 3, GlideTrack performs best on slope_track16 indicating the highest HOTA, IDF1, and MOTA scores. Examples of each viewpoint can be seen in Fig. 7. This sequence benefits from a clear viewpoint and moderate crowding that allows the trajectory model to maintain long and uninterrupted tracks. The two sequences captured from the same vantage slope_track16 and slope_track17 illustrate the impact of scene density. The heavier traffic in slope_track17 reduces the association accuracy and identity scores as skiers overlap and occlude one another making the trajectories more complex. slope_track18 is the most spatially complex by including multiple slopes with entrances and exits both near and far from the camera. This variety introduces scale changes and unpredictable motion which lowers HOTA and MOTA scores. slope_track19 highlights detector noise. Background objects are occasionally flagged as people increasing false positives and suppressing MOTA and DetA despite a solid association score. Finally, slope_track20 presents a wide entry field and closer camera placement. In this sequence, GlideTrack handles these perspective changes well but shows occasional identity switches for individuals standing nearest the lens.

Fig. 6. Comparison of challenging sequences where BoTSORT, OC-SORT, and HybridSORT lose track identities during long occlusions or abrupt skier motion, while GlideTrack maintains trajectories that closely follow the ground-truth annotations.

Table 3
Per-video evaluation of GlideTrack on the Slope-Track test set.

Video	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑	LocA↑
slope_track16	75.7	76.7	65.9	91.7	86.9	94.7
slope_track17	65.8	66.2	56.4	80.2	76.8	94.7
slope_track18	60.9	64.9	61.6	63.4	60.2	93.8
slope_track19	69.8	74.0	74.7	56.4	65.2	95.6
slope_track20	73.1	76.3	67.5	83.3	79.2	94.8
Combined	68.0	70.3	63.6	74.7	72.8	94.8

5.4.2. Analysis of the different normalization strategies

We discuss several normalization strategies (Table 4) to identify the most suitable for our use case. The Slope-Track dataset presents challenges such as varying resolutions, frame rates, viewpoints and slope direction. Using normalization helps reduce the amount of diverse data needed for the model to generalize effectively. We also decided on offsets instead of absolute positions because it provides stable and relative motion cues that focus on changes in position and size between frames.

Normalization strategy 1 compensates for sudden zoom or distance changes. However, it operates locally on a frame-to-frame basis and does not capture the overall dynamics of the skier. To address this, normalization strategy 2 introduces a temporal factor to add a velocity-like component. This helps distinguish skiers moving at different speeds but, it is sensitive to missed detections or irregular frame intervals which can occur on slopes.

For normalization strategies 3 and 4, the strong correlation between skier height in pixels and actual depth makes height-based normalization effective when the camera is positioned downhill. Using the trajectory's average height further stabilizes long downhill sequences. However, height alone becomes unreliable when the camera is not at the base of the slope, and the moving average requires enough observations to be effective.

An absolute image-size normalization (Normalization 5) operates at the frame level and is independent of object size. It is robust to

Fig. 7. Examples of frames from the five Slope-Track sequences. The top row shows slope_track16, slope_track17, and slope_track18, while the bottom row shows slope_track19 and slope_track20.

Table 4
Normalization strategies used in the Mamba-based trajectory model.

Normalization #	Normalization strategy
1	$\frac{x_c - x_1}{w_c}, \frac{y_c - y_1}{h_c}, \frac{h_c - h_1}{h_c}, \frac{w_c - w_1}{w_c}$
2	$\frac{x_c - x_1}{(t_c - t_1)w_c}, \frac{y_c - y_1}{(t_c - t_1)h_c}, \frac{h_c - h_1}{(t_c - t_1)h_c}, \frac{w_c - w_1}{(t_c - t_1)w_c}$
3	$\frac{x_c - x_1}{h_c}, \frac{y_c - y_1}{h_c}, \frac{h_c - h_1}{h_c}, \frac{w_c - w_1}{h_c}$
4	$\frac{x_c - x_1}{\frac{1}{c}\sum_{i=1}^{c} h_i}, \frac{y_c - y_1}{\frac{1}{c}\sum_{i=1}^{c} h_i}, \frac{h_c - h_1}{\frac{1}{c}\sum_{i=1}^{c} h_i}, \frac{w_c - w_1}{\frac{1}{c}\sum_{i=1}^{c} h_i}$
5	$\frac{x_c}{I_w}, \frac{y_c}{I_h}, \frac{w_c}{I_w}, \frac{h_c}{I_h}$
6	$\frac{x_c - x_1}{t_c - t_1}, \frac{y_c - y_1}{t_c - t_1}, \frac{h_c - h_1}{t_c - t_1}, \frac{w_c - w_1}{t_c - t_1}$
7	$\frac{x_c - x_1}{\frac{1}{c}\sum_{i=1}^{c} w_i}, \frac{y_c - y_1}{\frac{1}{c}\sum_{i=1}^{c} h_i}, \frac{h_c - h_1}{\frac{1}{c}\sum_{i=1}^{c} h_i}, \frac{w_c - w_1}{\frac{1}{c}\sum_{i=1}^{c} w_i}$

bounding-box noise but does not account for the natural changes in skier scale along the slope. As skiers approach the camera and their bounding boxes grow, the normalized values increase proportionally, making it harder for the model to extrapolate motion accurately.

Normalization 6 emphasizes raw motion in pixel space without considering object size. This highlights velocity cues and works well for side-view slopes where apparent size remains nearly constant. However, for frontal-downhill cameras, the lack of scale compensation makes it difficult to predict future positions.

Finally, normalization 7 uses track-level averages of both width and height to normalize positional and size changes. This provides stable scale information across frames by reducing sensitivity to variations in skier posture or orientation. Similar to normalization 6, it requires sufficient trajectory history but offers the added benefit of combining both width and height for better overall scale representation.

We opted to use normalization 7.

5.4.3. Analysis of the embeddings

We evaluate the contributions of the spatial embedding and the normalized bounding box (content) representation to the overall tracking performance. All experiments are conducted using the Deep-EIoU tracker without appearance features, integrating only the proposed embeddings in GlideTrack.

Table 5
Ablation study on spatial and content embeddings with varying spatial embedding dimension d.

Content	Spatial	d	HOTA↑	AssA↑	DetA↑	IDF1↑	MOTA↑
			65.0	58.4	72.4	66.7	74.2
	✓	2	65.5	59.2	72.4	66.4	74.3
	✓	4	66.4	60.8	72.6	67.8	74.4
	✓	8	66.4	60.7	72.6	67.9	74.4
	✓	16	65.4	59.0	72.5	66.4	74.4
✓			67.6	62.9	72.7	69.6	74.6
✓	✓	2	66.7	61.3	72.5	69.7	74.6
✓	✓	4	67.3	62.5	72.5	69.7	74.4
✓	✓	8	**68.0**	**63.6**	**72.8**	**70.3**	**74.7**
✓	✓	16	67.2	62.3	72.4	69.3	74.3

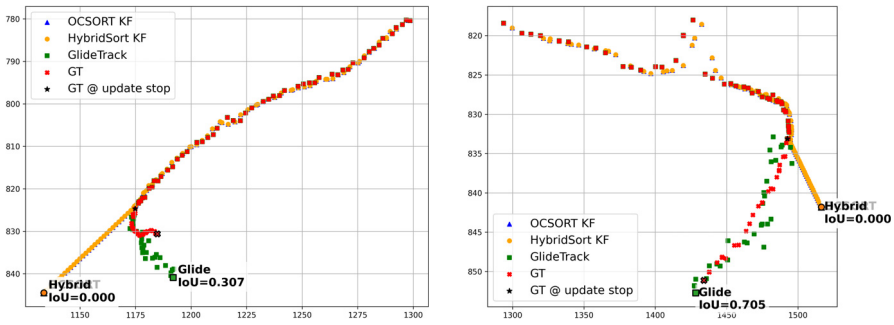
As shown in Table 5, both the spatial embedding and the normalized bounding box representation individually improve the tracking performance over the baseline algorithm. In particular, introducing the spatial embedding alone leads to consistent improvements in HOTA and AssA across different embedding dimensions. We observed that incorporating the content representation improves performance, especially in association-related metrics such as HOTA, AssA and IDF1.

Additionally, we analyzed the impact of the spatial embedding dimension d. While transformer-based architectures often rely on higher-dimensional embeddings to encode rich positional information, our experiments indicate that smaller embedding dimensions are more effective in this setting. We consider spatial embedding dimensions $d \in 2, 4, 8, 16$ using either the spatial embedding alone or in combination with the content representation. As shown in Table 5, increasing d does not result in consistent improvements across tracking metrics. For both configurations, the best performance achieved with $d = 8$, whereas larger values of d lead to slight performance degradation.

This behavior is likely due to both the limited size of the training dataset and the low dimensionality of the input. When the dimensionality of the sinusoidal embedding is increased, the representational capacity of the model is enlarged. In this context, this effect promotes overfitting rather than improved generalization.

5.4.4. Kalman Filter vs. GlideTrack under missing gaps

We validated GlideTrack's ability to handle non-linear motion by comparing its predicted trajectories with those of several Kalman filter-based trackers. As discussed in Section 3.2.4, the motion of skiers on varying slopes is highly non-linear, making it difficult to predict



(a) 30-frame prediction gap following 180 frames (b) 30-frame prediction gap following 180 frames

Fig. 8. Comparison of predicted trajectories in challenging non-linear motion segments (turns and speed changes). Each plot shows the center-x and center-y positions of the object as predicted by different trackers during forced prediction intervals. We additionally report the IoU between the final predicted position and the corresponding ground-truth box, providing a direct measure of spatial accuracy under occlusion.

when they are not seen. For each sequence, we provide ground-truth bounding boxes to a set of Kalman Filters (Cao et al., 2023; Yang et al., 2024) and to the GlideTrack model. After a fixed number of frames, we stop updating the Kalman filters and force them to predict for 30 frames to simulate intervals when an object becomes occluded or is not detected.

For Kalman-based trackers, we observed that they immediately revert to linear motion predictions once updates stop, as illustrated in Fig. 8. This behavior is expected because standard Kalman motion models assume constant velocity unless corrected by new observations. As a result, the predicted trajectory diverges from the true path almost immediately after the gap begins. In particular, when a skier becomes unobserved and accelerates, Kalman filter-based methods continue to predict positions based on the previous velocity, causing the predicted points to cluster too closely together, as shown in Fig. 8(b). Trackers such as HybridSORT (Yang et al., 2024) and OCSORT (Cao et al., 2023) augment their Kalman Filters with velocity perturbation to prevent perfectly linear predictions, but these adjustments still operate within a fundamentally linear framework. Consequently, they struggle to capture the non-linear accelerations and decelerations typical of ski slopes. In contrast, GlideTrack continues to produce nonlinear predictions that better reflect the object's underlying dynamics as seen in Fig. 8. This allows GlideTrack to follow complex motion patterns that linear models cannot reproduce.

5.5. Limitations

Despite its strong performance, GlideTrack still faces several practical constraints. Firstly, the trajectory model depends on a sufficiently long history. Therefore, tracks with only a few past observations such as newly initialized objects make it harder for the model to infer reliable motion patterns. Secondly, the prediction accuracy degrades when extrapolating more than about 60 frames into the future, as compounding errors grow.

6. Conclusion

In this paper, we presented SlopeTrack. Slope-Track is a new benchmark dataset for multi-object tracking in snow-based environments, which is designed to capture the complexities of real-world ski slope monitoring. Extensive benchmarking shows that standard trackers struggle on Slope-Track, particularly due to unreliable appearance features and the challenges of maintaining identity under occlusion and motion variability. To address this, we introduced the motion-driven trajectory model, GlideTrack. We showed that GlideTrack coupled with Deep-EIoU can outperform the Kalman-based methods and learned-motion baselines on SlopeTrack to demonstrate the value of learned sequence to sequence modeling in this domain. While Slope-Track still has limitations, it establishes a much-needed benchmark for unconstrained outdoor and sports tracking, and our results with GlideTrack highlight promising directions for future research.

CRediT authorship contribution statement

M'Saydez Campbell: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Christophe Ducottet:** Writing – review & editing, Validation, Supervision, Conceptualization. **Damien Muselet:** Writing – review & editing, Supervision, Conceptualization. **Rémi Emonet:** Writing – review & editing, Validation, Supervision, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is part of the collaborative i-Démo Régionalisé SOFTEN project (Solution pour l'Optimisation des Flux de skieurs pour la Transition ENergétique des domaines skiables) of the French government's regionalized France 2030 program. It was funded by BPI France, Direction Générale de la Compétitivité de l'Industrie et des Services (DGCIS), Grenoble Alpes Métropole and the Auvergne-Rhône-Alpes Region.

Data availability

The link to the code and dataset is available in the abstract.

References

Aharon, N., Orfaig, R., Bobrovsky, B.-Z., 2022. Bot-SORT: Robust associations multi-pedestrian tracking. arXiv preprint arXiv:2206.14651.

Akyon, F.C., Altinuc, S.O., Temizel, A., 2022. Slicing Aided Hyper Inference and Fine-tuning for Small Object Detection. In: 2022 IEEE Int. Conf. Image Process.. ICIP, pp. 966–970. http://dx.doi.org/10.1109/ICIP46576.2022.9897990.

Akyon, F.C., Cengiz, C., Altinuc, S.O., Cavusoglu, D., Sahin, K., Eryuksel, O., 2021. SAHI: A lightweight vision library for performing large scale object detection and instance segmentation. http://dx.doi.org/10.5281/zenodo.5718950, [Software].

Association, N.S.A., 2024. Historical Skier Visits: 1978/79-2023/24. URL https://nsaa.org/webdocs/Media_Public/IndustryStats/Historical_Skier_Visits_2024.pdf.

Bachmann, R., Spörri, J., Fua, P., Rhodin, H., 2019. Motion capture from pan-tilt cameras with unknown orientation. In: 2019 Int. Conf. on 3D Vis. (3DV). pp. 308–317. http://dx.doi.org/10.1109/3DV.2019.00042, URL https://doi.ieeecomputersociety.org/10.1109/3DV.2019.00042.

Bernardin, K., Stiefelhagen, R., 2008. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. EURASIP J. Image Video Process. 1–31. http://dx.doi.org/10.1155/2008/246309, URL https://doi.org/10.1155/2008/246309.

Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016. Simple online and realtime tracking. In: 2016 IEEE Int. Conf. on Image Process.. ICIP, IEEE, pp. 3464–3468. http://dx.doi.org/10.1109/icip.2016.7533003.

Cao, J., Pang, J., Weng, X., Khirodkar, R., Kitani, K., 2023. Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking. In: Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.. pp. 9686–9696. http://dx.doi.org/10.1109/CVPR52729.2023.00934.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: Comput. Vis. – ECCV 2020: 16th European Conf., Glasgow, UK, August 23–28, 2020, Proceed., Part I. Springer-Verlag, Berlin, Heidelberg, pp. 213–229. http://dx.doi.org/10.1007/978-3-030-58452-8_13, URL https://doi.org/10.1007/978-3-030-58452-8_13.

Chen, G., Wang, H., Chen, K., Li, Z., Song, Z., Liu, Y., Chen, W., Knoll, A., 2022. A Survey of the Four Pillars for Small Object Detection: Multiscale Representation, Contextual Information, Super-Resolution, and Region Proposal. IEEE Trans. Syst, Man, Cybern. : Syst. 52 (2), 936–953. http://dx.doi.org/10.1109/TSMC.2020.3005231.

Cioppa, A., Giancola, S., Deliège, A., Kang, L., Zhou, X., Cheng, Z., Ghanem, B., Van Droogenbroeck, M., 2022. SoccerNet-Tracking: Multiple Object Tracking Dataset and Benchmark in Soccer Videos. In: Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR) Workshops. pp. 3491–3502. http://dx.doi.org/10.1109/CVPRW56347.2022.00393, URL https://doi.ieeecomputersociety.org/10.1109/CVPRW56347.2022.00393.

Corporation, C., 2024. Computer vision annotation tool (CVAT). http://dx.doi.org/10.5281/zenodo.12771595, [Software].

Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L., 2023. SportsMOT: A Large Multi-Object Tracking Dataset in Multiple Sports Scenes. In: Proc. of the IEEE/CVF Int. Conf. on Comput. Vis.. ICCV, pp. 9921–9931. http://dx.doi.org/10.1109/ICCV51070.2023.00910, URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00910.

Dao, T., Gu, A., 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In: International Conference on Machine Learning. ICML.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L., 2020. MOT20: A benchmark for multi object tracking in crowded scenes. ArXiv:2003.09003.

Ding, J., Xue, N., Xia, G.-S., Bai, X., Yang, W., Yang, M.Y., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2022. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. IEEE Trans. Pattern Anal. Mach. Intell. 44 (11), 7778–7796. http://dx.doi.org/10.1109/TPAMI.2021.3117983.

Dunnhofer, M., Sordi, L., Martinel, N., Micheloni, C., 2024. Tracking Skiers From the Top to the Bottom. In: Proc. of the IEEE/CVF Winter Conf. on Appl. of Comput. Vis.. WACV, pp. 8511–8521. http://dx.doi.org/10.1109/WACV57701.2024.00832, URL https://doi.ieeecomputersociety.org/10.1109/WACV57701.2024.00832.

Gao, R., Wang, L., 2023. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV, pp. 9901–9910.

Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., 2021. YOLOX: Exceeding YOLO series in 2021. ArXiv:2107.08430.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conf. on Comput. Vis. and Pattern Recognit.. pp. 3354–3361. http://dx.doi.org/10.1109/CVPR.2012.6248074.

Gu, A., Dao, T., 2023. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

Hu, B., Luo, R., Liu, Z., Wang, C., Liu, W., 2024. Trackssm: A general motion predictor by state-space model. URL https://arxiv.org/abs/2409.00487. arXiv:2409.00487.

Huang, H.-W., Yang, C.-Y., Chai, W., Jiang, Z., Hwang, J.-N., 2025. Mambamot: State-space model as motion predictor for multi-object tracking. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5. http://dx.doi.org/10.1109/ICASSP49660.2025.10890199.

Huang, H.-W., Yang, C.-Y., Sun, J., Kim, P.-K., Kim, K.-J., Lee, K., Huang, C.-I., Hwang, J.-N., 2024. Iterative Scale-Up ExpansionIoU and Deep Features Association for Multi-Object Tracking in Sports. In: Proc. of the IEEE/CVF Winter Conf. on Appl. of Comput. Vis.. pp. 163–172. http://dx.doi.org/10.1109/WACVW60836.2024.00024, URL https://doi.ieeecomputersociety.org/10.1109/WACVW60836.2024.00024.

Jocher, G., Qiu, J., Chaurasia, A., 2023. Ultralytics YOLO. [Software]. URL https://github.com/ultralytics/ultralytics.

Ludwig, K., Lorenz, J., Schön, R., Lienhart, R., 2023. All Keypoints You Need: Detecting Arbitrary Keypoints on the Body of Triple, High, and Long Jump Athletes. In: Proc. of the 2023 IEEE/CVF Int. Conf. on Comput. Vis. and Pattern Recognit. Workshops. CVPRW, pp. 5179–5187. http://dx.doi.org/10.1109/CVPRW59228.2023.00546, URL https://doi.ieeecomputersociety.org/10.1109/CVPRW59228.2023.00546.

Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B., 2020. HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking. Int. J. Comput. Vis. 1–31. http://dx.doi.org/10.1007/s11263-020-01375-2, URL https://doi.org/10.1007/s11263-020-01375-2.

Magazine, S., 2023a. 108 web cams are a popular page on any resort site. URL https://www.saminfo.com/page/13066-108-web-cams-are-a-popular-page-on-any-resort-site-but-is-the-widening-gap-an-issue. [Accessed: 29 September 2025].

Magazine, S.O., 2023b. Artificial intelligence for ski resorts. URL https://www.snowopsmag.com/article/artificial-intelligence-for-ski-resorts/. [Accessed: 29 September 2025].

Maggiolino, G., Ahmad, A., Cao, J., Kitani, K., 2023. Deep OC-sort: Multi-pedestrian tracking by adaptive re-identification. In: 2023 IEEE Int. Conference on Image Process.. ICIP, pp. 3025–3029. http://dx.doi.org/10.1109/ICIP49359.2023.10222576.

Magnusson, T., Frei, A., Rixen, R., Jonas, P., 2020. Towards a webcam-based snow cover monitoring network: Methodology and evaluation. Cryosphere 14, 1409–1427. http://dx.doi.org/10.5194/tc-14-1409-2020, URL https://tc.copernicus.org/articles/14/1409/2020/. [Accessed: 29 September 2025].

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K., 2016. MOT16: A benchmark for multi-object tracking. ArXiv:1603.00831 [Cs].

Pei, Z., 2019. Deepsort pytorch. Pytorch implentation of Deep-SORT [Software]. URL https://github.com/ZQPei/deep_sort_pytorch.

Qin, Z., Wang, L., Zhou, S., Fu, P., Hua, G., Tang, W., 2024. Towards Generalizable Multi-Object Tracking. In: Proceedings of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.. CVPR, pp. 18995–19004. http://dx.doi.org/10.1109/CVPR52733.2024.01797.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C., 2016. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. In: Hua, G., Jégou, H. (Eds.), Comput. Vis. – ECCV 2016 Workshops. Springer International Publishing, Cham, pp. 17–35. http://dx.doi.org/10.1007/978-3-319-48881-3{_}2, URL https://doi.org/10.1007/978-3-319-48881-3_2.

Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P., 2022. DanceTrack: Multi-object tracking in uniform appearance and diverse motion. In: Proc. of the IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit.. CVPR, pp. 20993–21002. http://dx.doi.org/10.1109/CVPR52688.2022.02032, URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.02032.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), In: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.-S., Bai, X., 2019. iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images. In: Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recognit. Workshops. pp. 28–37.

Xiao, C., Cao, Q., Luo, Z., Lan, L., 2024. MambaTrack: A simple baseline for multiple object tracking with state space model. In: Proceedings of the 32nd ACM International Conference on Multimedia. MM '24, Association for Computing Machinery, New York, NY, USA, pp. 4082–4091. http://dx.doi.org/10.1145/3664647.3680944, URL https://doi.org/10.1145/3664647.3680944.

Yan, F., Luo, W., Zhong, Y., Gan, Y., Ma, L., 2023. Bridging the gap between end-to-end and non-end-to-end multi-object tracking. URL https://arxiv.org/abs/2305.12724. arXiv:2305.12724.

Yan, B., Peng, H., Fu, J., Wang, D., Lu, H., 2021. Learning spatio-temporal transformer for visual tracking. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 10428–10437. http://dx.doi.org/10.1109/ICCV48922.2021.01028.

Yang, M., Han, G., Yan, B., Zhang, W., Qi, J., Lu, H., Wang, D., 2024. Hybrid-SORT: Weak cues matter for online multi-object tracking. In: Proc. of the AAAI Conf, on Artif. Intell.. vol. 38, pp. 6504–6512. http://dx.doi.org/10.1609/aaai.v38i7.28471, URL https://doi.org/10.1609/aaai.v38i7.28471.

Yi, K., Luo, K., Luo, X., Huang, J., Wu, H., Hu, R., Hao, W., 2024. UCMCTrack: Multi-Object Tracking with Uniform Camera Motion Compensation. Proc. the AAAI Conf. Artif. Intell. 38 (7), 6702–6710. http://dx.doi.org/10.1609/aaai.v38i7.28493, URL https://doi.org/10.1609/aaai.v38i7.28493.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X., 2022. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. In: Proc. of the European Conf. on Comput. Vision. ECCV, pp. 1–21. http://dx.doi.org/10.1007/978-3-031-20047-2{_}1, URL https://doi.org/10.1007/978-3-031-20047-2_1.

Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H., 2021. Detection and Tracking Meet Drones Challenge. IEEE Trans. Pattern Anal. Mach. Intell. http://dx.doi.org/10.1109/TPAMI.2021.3119563, URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3119563.

Zwölfer, M., Heinrich, D., Schindelwig, K., Wandt, B., Rhodin, H., Spörri, J., Nachbauer, W., 2021. Improved 2D Keypoint Detection in Out-of-Balance and Fall Situations - combining input rotations and a kinematic model. CoRR arXiv:2112.12193.

Zwölfer, M., Heinrich, D., Schindelwig, K., Wandt, B., Rhodin, H., Spörri, J., Nachbauer, W., 2023. Deep learning-based 2D keypoint detection in alpine ski racing – A performance analysis of state-of-the-art algorithms applied to regular skiing and injury situations. JSAMS Plus (ISSN: 2772-6967) 2, 100034. http://dx.doi.org/10.1016/j.jsampl.2023.100034, URL https://doi.org/10.1016/j.jsampl.2023.100034.